**Universidad Autónoma de Madrid**

**Departamento de Biología**
**Facultad de Ciencias**

**Consejo Superior de Investigaciones Científicas**

**Departamento de Biodiversidad y Biología Evolutiva**
**Museo Nacional de Ciencias Naturales**

# MODELOS PREDICTIVOS DE DIVERSIDAD: APLICACIÓN AL ESTUDIO DE LA RIQUEZA DE ARANEIDOS Y TOMÍSIDOS (ARACHNIDA, ARANEAE, ARANEIDAE & THOMISIDAE) EN LA COMUNIDAD DE MADRID Y A LA PREDICCIÓN DE LA DISTRIBUCIÓN DE *MACROTHELE CALPEIANA* (ARACHNIDA, ARANEAE, HEXATHELIDAE) EN LA PENÍNSULA IBÉRICA

*Memoria presentada por* ALBERTO JIMÉNEZ VALVERDE *para optar al Grado de Doctor en Ciencias Biológicas*

Alberto Jiménez Valverde

Vº Bº Director de Tesis

Dr. Jorge Miguel Lobo

Vº Bº Tutor de Tesis

Dr. José Martín Cano

Madrid, Junio de 2006

*A mis padres, por todo*

# AGRADECIMIENTOS

Jorge confió en mí desde el primer momento, dándome todo su apoyo siempre que lo he necesitado, día a día durante estos cinco años. Ha sido maestro, el mejor, pero sobre todo ha sido compañero, colega, amigo. Jorge, has hecho que los días en el Museo pasaran rápido, que disfrutara enormemente con nuestro trabajo. Me viene a la cabeza una conversación que tuvimos un día, no era sobre ciencia, era una de esas en las que se dicen cosas importantes. A lo largo de la vida se conoce a mucha gente. Unos pasan y otros se quedan. Entre todos, y no siempre ocurre, aparece alguien que brilla de una manera especial, alguien del que te sientes verdaderamente afortunado por haber conocido. Tú eres una de esas personas. No cambies nunca.

Qué puedo decir de Joaquín. A su lado he aprendido mucho durante estos cinco años, acelerando mi aprendizaje en cuestiones de SIG, estadística y ecología. Su huella está presente en cada página de esta tesis. Pero, ante todo, él me ha dejado claro en qué consiste el compañerismo: es contar, es escuchar, es compartir, es dar sin esperar nada a cambio. Joaquín, eres un tío grande.

Vicente ha sido un gran compañero de viaje. Gracias por transmitirme tu desbordante pasión por el mundo que nos rodea. Gracias por tus consejos. Gracias por creer en mí, y gracias por no cansarte de repetírmelo.

Andrés y Jose han sido testigos de todo el proceso de creación de la tesis. Gracias por todo lo que hemos compartido juntos, vuestra amistad ha sido el mejor regalo. He de mencionar también al resto de compañeros del Museo: Paco, Javi, Tere, Pili…y todos aquellos a los que me dejo.

La Comunidad de Madrid me concedió la beca Museo Nacional de Ciencias Naturales/CSIC/Comunidad de Madrid, gracias a la cual he podido realizar esta tesis

doctoral. La Consejería de Medio Ambiente de Madrid concedió, año tras año, los permisos necesarios para efectuar las colectas de arañas.

Gracias a mis padres hoy esta tesis ve la luz. Vosotros me enseñasteis a luchar por mis sueños, a fijar un objetivo y a no desfallecer en el camino hasta conseguirlo. Es la única manera de ser feliz, me decíais. Desde pequeño incentivasteis mi afición por la Naturaleza y nunca escuche una palabra que intentase hacerme cambiar de idea; siempre quise dedicarme a la biología y hoy estoy aquí, impulsado por vuestros ánimos. A vosotros os lo debo todo, mil gracias.

Laura, has tenido que aguantar al pesado de tu hermano, lo sé. Y lo que te queda. Qué hubiera sido de mí todo este tiempo sin las risas que nos echamos, que han hecho más llevaderos los largos días frente al ordenador. Mi mayor fortuna es tenerte como hermana.

Mis tías me han apoyado igualmente sin condiciones. Sensi, tú incluso le diste al azadón para colocar trampas en una de las zonas más áridas de Madrid. Recuerdo el calor sofocante. Es una suerte tener una familia como vosotras.

Con mi amiga Marisa he compartido los momentos felices y los más duros desde los días de la UAM. Saber que estabas ahí, Marisa, ha sido un motor cada mañana. No dejes nunca de sonreír.

Con Raquel viví los inicios del camino, gracias por esos momentos. Sé que me dejo a mucha gente sin mencionar. Javi, Cristina, Celia, David, Carlos, Eva…, muchas gracias a todos. De corazón.

*Alberto*

# ÍNDICE

# Introducción

El número de especies actualmente conocido es inferior a 2 millones, mientras que las estimas sobre el número que quedan aún por descubrir varían entre 5 y 10 millones (Groombridge & Jenkins, 2002). Cada año son descritas nuevas especies, incluso en taxones relativamente bien conocidos como mamíferos y aves (Diamond, 1985; Medellín & Soberón, 1999; Patterson, 1994, 2000), aunque son los artrópodos, con alrededor de 1 millón de especies descritas (Groombridge & Jenkins, 2002), el grupo que mayor número de especies nuevas aporta. Aunque el número de taxónomos trabajando con vertebrados, invertebrados y plantas vasculares es similar, considerando las diferencias en diversidad entre los tres grupos, la tasa de esfuerzo taxonómico esta claramente sesgada hacia los grupos menos diversos pero más llamativos y fáciles de estudiar (May, 1988, 1994). De igual manera, se estima que tan solo el 6% de los taxónomos desarrolla su actividad en países en vías de desarrollo (Gaston & May, 1992). Este sesgo geográfico implica que la mayor parte del esfuerzo taxonómico se desarrolla en las áreas de menor diversidad. Pero incluso a menor escala, en regiones y grupos concretos, los sesgos geográficos son también llamativos y evidentes (por ejemplo, Cabrero-Sañudo & Lobo, 2003; Baselga *et al.*, en prensa; Jiménez-Valverde & Ortuño, en prensa).

No solo nos enfrentamos al desconocimiento del número total de especies existentes, sino que, de las que conocemos, disponemos de pocos datos, la mayoría de ellos segados, sobre su distribución (Graham *et al*., 2004). En general, son las áreas más atractivas o de más fácil acceso las que reciben mayor número de visitas por parte de los investigadores (Dennis *et al*., 1999; Dennis & Thomas, 2000; García-Barros *et al*., 2000; Reddy & Dávalos, 2003; Reutter *et al*., 2003; Richardson *et al*., 2006), por lo que las áreas de distribución de los organismos son visiones parciales que reflejan

un patrón espacial histórico naturalista más que un patrón de distribución real. Por tanto, ignoramos los patrones geográficos de la vida y somos incapaces de describir la distribución de la mayoría de las especies, y de definir tanto el número de taxones como su identidad para cualquier localidad del globo. Dada la dimensión de la biodiversidad y las limitaciones humanas, el desconocimiento general que tenemos de su distribución y magnitud global probablemente sea una limitación insalvable e inevitable.

Por otra parte, la disminución de la variedad de la vida y la extinción de especies son actualmente hechos constatados (Pimm *et al*., 1995; Lawton & May, 1995; Chapin *et al*., 2000; Pimm & Raven, 2000). El ritmo de extinción masiva es tan elevado que entre los científicos existe la opinión generalizada de que nos encontramos ante una verdadera crisis de la biodiversidad (Chapin *et al*., 2000). Actualmente, la tasa de extinción de especies es mucho mayor que la tasa de especiación (Wilson, 1992) y, en algunos grupos, es probable que ya supere la tasa de descubrimiento de nuevos taxones (Diamond, 1985). Las estimaciones más fiables de esta tasa de extinción rondan la media de 40.000 especies/año (entre 27.000 y 250.000; Lomborg 2001), lo que nos llevaría a perder la práctica totalidad de las especies en unos pocos siglos (Stork, 1997). Aunque estas estimas puedan ser exageradas (Lomborg, 2001), la crisis de la biodiversidad es un hecho palpable y constatado. La situación es especialmente preocupante para los denominados *grupos hiperdiversos* ya que, dada su desbordante variedad, no es previsible que se alcance el conocimiento necesario para su protección en un tiempo razonable (Hammond, 1994; Dunn, 2005).

Sin embargo, conocer la distribución geográfica de las diversidad biológica es una necesidad imperiosa para abordar cuestiones de muy diversa índole, relacionadas con

la ecología, la biogeografía y la evolución (Guisan & Thuiller, 2005; Samways, 2005). Además, para llevar a cabo programas de conservación sólidos desde un punto de vista científico, es imprescindible conocer tanto la distribución de las diferentes especies (por ejemplo, Dobson *et al*., 1997; van Jaarsveld *et al*., 1998; Howard *et al*., 1998; Araújo, 1999; Araújo & Williams, 2000; Andriamampianina *et al.*, 2000; Polasky *et al*, 2000; Martín-Piera, 2001) como los diferentes atributos de la biodiversidad para un conjunto de localidades dado, con el fin de proteger lugares especialmente ricos en especies o de gran rareza o endemicidad (por ejemplo, Margules *et al*., 1988; Zimmermann & Kienast, 1999; Ferrier *et al*., 2002a; Gladstone, 2002; Araújo *et al*., 2004). Además, es necesario conocer las causas de esos patrones observados con el fin de abordar la protección de la naturaleza bajo una base sólida de comprensión de los fenómenos. Pero, considerando nuestra más que probable insalvable ignorancia, ¿cómo podemos abordar el reto de comprender lo desconocido?

## BUSCANDO UN ATAJO: LOS MODELOS PREDICTIVOS

Los modelos predictivos son técnicas empleadas tanto para interpolar como para extrapolar patrones naturales a territorios carentes de información. Básicamente, son funciones que relacionan el atributo de interés con una serie de variables explicativas (ver Fig. 1; Nicholls, 1989; Ferrier, 2002; Ferrier *et al.*, 2002a, b). El desarrollo de los Sistemas de Información Geográfica (SIG) ha posibilitado el almacenamiento, manejo y análisis cuantitativo de grandes cantidades de datos espaciales (Johnston, 1998). De esta manera, se puede disponer de información ambiental para cada localidad que cuente con información biológica. Estas variables se pueden someter a diversos análisis estadísticos con el fin de formalizar la relación variable biológica-ambiente en

un modelo matemático. Disponiendo de las capas temáticas digitales para un área determinada, el modelo resultante puede ser interpolado y/o extrapolado al territorio en cuestión generando un mapa predictivo de la variable de interés. Además, las funciones generadas y los patrones de ellas derivados pueden ser interpretados de cara a comprender los procesos subyacentes.

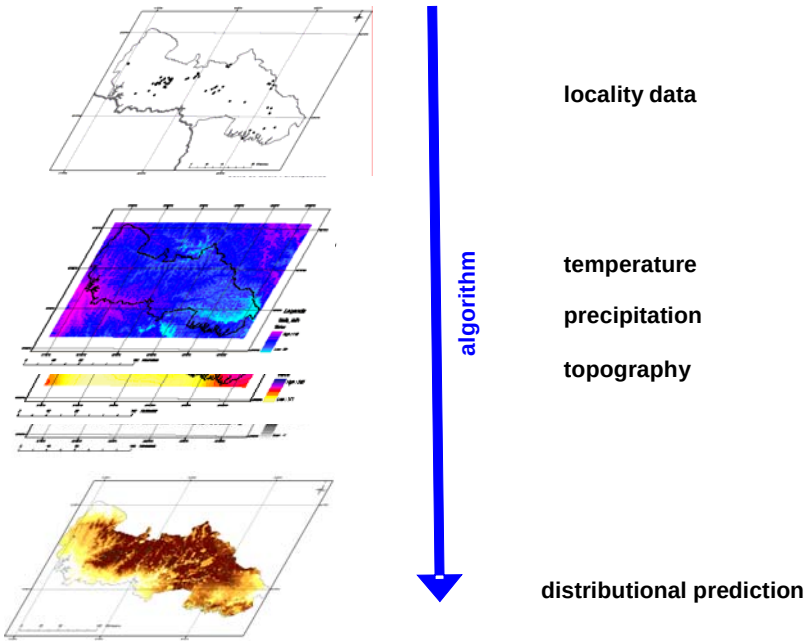



Figura 1.- Proceso de modelización de la distribución potencial de una especie empleando, como factores explicativos, variables ambientales almacenadas en un Sistema de Información Geográfica.

**Aproximación autoecológica. —** Los modelos predictivos se aplican a datos de distribución de especies concretas con el fin de predecir sus rangos geográficos (Guisan & Zimmermann, 2000; Scott *et al*., 2002; Guisan & Thuiller, 2005). Esta información pueden consistir en datos de presencia-ausencia o en datos de abundancia, aunque ésta última ha demostrado ser bastante más difícil de modelizar (por ejemplo, Pearce & Ferrier, 2001; Nielsen *et al*., 2005).

**Aproximación sinecológica. —** Consiste en modelizar los diferentes atributos de la biodiversidad (riqueza específica, rareza, endemicidad, etc.; Lobo & Martín-

Piera, 2002; Lobo *et al*., 2001, 2004; Hortal *et al*., 2001, 2003, 2004). Aunque son varios los atributos, la práctica totalidad de los trabajos se centran exclusivamente en modelizar la riqueza de especies.

## APROXIMACIÓN SINECOLÓGICA: MODELOS PREDICTIVOS DE LOS ATRIBUTOS DE LA BIODIVERSIDAD

Dado que el número de especies suele estar correlacionado con otras medidas de diversidad (Gaston, 1996), el estudio de los patrones espaciales y las causas de las variaciones en la riqueza específica ha sido uno de los temas centrales en Ecología (Huston, 1994; Miller, 1994). Así, los trabajos existentes que abordan la identificación de patrones y que tratan de explicar sus causas son numerosos y sería inmensa la lista de referencias que podría citarse. En general, la metodología para modelizar atributos de la biodiversidad es bastante sólida, basándose, principalmente, en análisis de regresión (Lobo, 2000; Hortal & Lobo, 2001; Lobo *et al*., 2004). El primer y principal problema que surge cuando se desea abordar uno de estos estudios es contar con una buena variable dependiente. Ya hemos visto como los sesgos presentes en la información biológica implican que no contemos con inventarios fiables de la mayoría de las localidades y que, si estos existen, presentan un sesgo geográfico importante. Es la principal cuestión que hemos abordado en esta tesis antes de realizar el estudio sobre la riqueza de arañas en Madrid.

**La variable dependiente: inventarios fiables. —** Modelizar la riqueza específica (o cualquier otro atributo de la biodiversidad) implica que los inventarios que van a ser empleados para parametrizar el modelo deben de ser completos (Harper & Hawksworth, 1995). Es decir, todas las especies presentes en esos ensamblajes

tienen que haber sido registradas. Sin embargo, conviene puntualizar que el tamaño y la composición de un inventario de especies en un lugar determinado varía con el tiempo (ver Adler & Lauenroth, 2003) debido a una característica fundamental de la distribución espacial de las especies: sus rangos de distribución no son temporalmente estables. Una especie puede ampliar o reducir su distribución en función de cambios en el ambiente. Además, determinadas especies pueden variar su fenología en función, por ejemplo, de las condiciones de un año determinado, pudiendo llegar a no emerger o ser detectables todos los años. Por otra parte, los individuos errantes (*vagrants*) son una fuente importante de sesgo en los inventarios (ver, por ejemplo, Dennis, 2001), ya que no pueden considerarse habitantes estrictos del área muestreada. Deben ser, no obstante, elementos importantes de la biodiversidad del lugar, ya que son responsables de parte de la resiliencia (capacidad de recuperación) de los ecosistemas frente a variaciones en las condiciones ambientales. La importancia de las variaciones temporales de la riqueza de especies y de los *vagrants* dependerá de la escala espacial de trabajo y de las características espacio-ambientales del área de trabajo (grado de heterogeneidad ambiental, de aislamiento, de hostilidad ambiental frente al grupo taxonómico de estudio, etapa sucesional, etc.). Por tanto, conviene tener presente que un inventario real no llega a completarse nunca, por lo que la estima final del número de especies depende de la resolución temporal y espacial que empleemos en el muestreo; es fundamental que las estimas de riqueza especifiquen el área y periodo temporal de recogida de muestras (Adler & Lauenroth, 2003).

Teniendo en cuenta estas consideraciones, la estimación de la fiabilidad de los inventarios es el primer paso necesario antes de elaborar cualquier modelo de riqueza (Hortal & Lobo, 2002). Tanto si se trabaja con datos recopilados a partir de fuentes

heterogéneas (bibliografía, colecciones, etc.) como si se obtienen inventarios a partir de muestreos de campo, una buena metodología para conocer su fiabilidad se basa en el empleo de curvas de acumulación, en las que la incorporación de nuevas especies al inventario se relaciona con alguna medida del esfuerzo de muestreo (ver Jiménez-Valverde & Hortal, 2003). Las curvas de acumulación permiten: 1) ofrecer una medida de confianza en los inventarios biológicos y posibilitar su comparación, 2) una mejor planificación del trabajo de muestreo, tras estimar el esfuerzo requerido para conseguir inventarios fiables, y 3) extrapolar el número de especies observado en un inventario para estimar el total de especies que estarían presentes en la zona (Soberón & Llorente, 1993; Colwell & Coddington, 1994; Gotelli & Colwell, 2001). La estimación del número de especies también puede realizarse empleando estimadores no paramétricos, basados en las técnicas de estimación del número de clases a partir de muestras y de captura-recaptura (Bunge & Fitzpatrick, 1993).

Cuando la información corológica ya disponible es insuficiente para obtener unos pocos inventarios con los que realizar un modelo (como es el caso de las arañas en la Península Ibérica) y nos vemos, por tanto, obligados realizar un muestreo de campo, es imprescindible conocer: i) el número y características de las técnicas de muestreo necesarias, ii) la cantidad de esfuerzo de muestreo a realizar, iii) si tanto las técnicas como el esfuerzo deben variar según las características ambientales de las parcelas de muestreo, y iv) el periodo temporal en el que llevar a cabo el muestreo a fin de obtener representaciones comparables de los ensamblajes. Estos aspectos, entre otros, han sido estudiados en los **capítulos 1 y 2** de la presente tesis, los cuales tratan de elaborar un protocolo de muestreo que garantice la obtención de inventarios fiables de las familias de arañas Araneidae y Thomisidae.

**Diseño espacial del muestreo. —** La localización de los puntos de muestreo es un proceso sumamente importante, tanto más cuanto mayor sea la relación entre el área de estudio y el número de localidades que es posible muestrear. Así, la situación de los puntos no debe presentar sesgos espaciales a fin de poder realizar inferencias fiable para la totalidad del área de interés (Yoccoz *et al*., 2001). Sin embargo, no solo basta con tener una buena representación espacial, sino que también se ha de conseguir representar lo mejor posible el gradiente ambiental (Hortal & Lobo, 2005). Una buena estrategia de selección debe, además, tener en cuenta el esfuerzo que es posible invertir en el muestreo de manera que, una vez fijado el número de localidades que es posible inventariar, la selección se realize maximizando el grado de representación espacio-ambiental en función del esfuerzo.

Todas estas características han sido tenidas en cuenta en el protocolo desarrollado por Hortal & Lobo (2005), en el cual se aplica al algoritmo *p-median* para seleccionar puntos a partir de una matriz de distancias espacio-ambiental. Nos fue imposible aplicar este protocolo a nuestros datos debido a impedimentos computacionales, por lo que nos vimos obligados a diseñar un método para seleccionar las localidades de muestreo que tuviera en cuenta todos los requerimientos expuestos. Este método, basado en un análisis de agrupamiento empleando el algoritmo *k-means*, se desarrolla en el **capítulo 3**.

**Validación. —** La validación de los modelos es un paso indispensable para evaluar su fiabilidad. Es más, lo procesos de validación basados en los mismos datos usados para entrenar los modelos deben ser evitados ya que proporcionan estimas demasiado optimistas de los errores (Chatfield, 1995; Olden & Jackson, 2000, Olden *et al*., 2002). Sin embargo, en la mayoría de los casos no se dispone de un conjunto de datos independientes con los que evaluar los modelos, ni resulta sencillo recabar esa

información. Es más, en caso de que se dispusiera de ella, ¿no sería mejor usar dicha información para realizar el modelo aumentando el tamaño de muestra? Debido a estos inconvenientes, numerosos estudios han empleado técnicas de partición de datos, en las que el conjunto de datos disponible se divide en dos grupos, uno para "entrenar" o realizar el modelo y otro para validar las predicciones obtenidas. La partición se puede realizar varias veces ("*k* fold partitioning") o tan sólo una única vez (*k*=2) (Fielding & Bell, 1997). Como ya hemos dicho, esta técnicas tienen el inconveniente de que reducen el tamaño muestral usado en la parmetrización de los modelos, con lo que sobreestiman el error predictivo (Fielding & Bell, 1997). Además, estos métodos son poco recomendables en el caso de contar con un bajo número de observaciones. El método de *jackknife*, recomendado por Olden & Jackson (2000) y Olden *et al*., (2002), evita estos problemas al extraer una sola observación en cada ocasión, elaborando el modelo con las *n*-1 observaciones restantes, validando cada modelo con la observación extraída y repitiendo el proceso *n* veces. Es el proceso de validación empleado en el **capítulo 4** (y en los **capítulos 9 y 10** de la parte autoecológica).

**Importancia relativa de los grupos de variables. —** Los modelos predictivos obtenidos a partir de métodos de regresión y mediante procedimientos automáticos de selección de variables no permiten analizar correctamente la importancia relativa de cada factor o grupo de factores debido a los problemas de multicolinealidad existente entre ellos (Mac Nally, 2000; Olden & Jackson, 2000). Sin embargo, muchas veces en el proceso de modelización, además de predecir, hay un interés por dilucidar el poder explicativo de los predictores. La partición jerarquizada (Mac Nally, 2002) y la partición de la varianza (Legendre & Legendre, 1998) permiten descomponer la variación presente en la variable dependiente en partes

debidas a los efectos independientes y combinados de cada factor o grupo de factores. Mediante estas técnicas no se generan modelos predictivos, son análisis complementarios si el deseo es ahondar en los posibles efectos causales (Mac Nally, 2002). En el **capítulo 4** se muestra la aplicación de la partición de la varianza (y en el **capítulo 10** se emplea, además, la partición jerarquizada en el estudio de la distribución de una especie).

**El modelo: la riqueza de araneidos y tomísidos en la Comunidad de Madrid. —** Todas estas cuestiones se han aplicado al estudio del patrón de riqueza específica de las familias de arañas Araneidae y Thomisidae en la Comunidad de Madrid y los factores que lo determinan (**capítulo 4**). En la Península Ibérica hay citadas, hasta el momento, 1210 especies de arañas, aunque indudablemente su número crecerá a medida que aumenten los estudios taxonómicos y faunísticos. La familia Araneidae cuenta con 56 especies repartidas en 22 géneros, mientras que la familia Thomisidae tiene 66 especies en 14 géneros. Los araneidos (Fig. 2) son especies, por lo general, de apariencia llamativa, constructoras de telas orbiculares que suelen anclar sobre la vegetación. Por el contrario, los tomísidos (Fig. 2) suelen ser especies crípticas que se camuflan en la vegetación o entre la hojarasca a la espera de las presas. Se eligieron estos dos grupos de arañas para estudiar los patrones de riqueza porque son, seguramente, las dos familias mejor conocidas desde el punto de vista taxonómico en la Península Ibérica, son relativamente fáciles de identificar y caen en abundancia en los muestreos.

La ausencia de tradición aracnológica en la Península Ibérica ha provocado que el conocimiento actual de su aracnofauna sea bastante limitado; los catálogos son escasos y la mayoría de las citas antiguas, muchas de ellas erróneas o, cuanto menos, dudosas, tal y como manifiestan Melic (2001) y Morano (2004). Además, la ausencia

y dispersión de la bibliografía necesaria para efectuar correctas identificaciones no facilita la labor de estudio. La Comunidad de Madrid no es ajena a este desconocimiento general, contando únicamente con un catálogo actualizado de la fauna de la familia Salticidae (Jiménez-Valverde, 2005). El trabajo realizado en la presente tesis doctoral ha permitido, además de elaborar una aproximación a las técnicas de muestreo y a los factores que determinan la riqueza de los araneidos y tomísidos, aumentar significativamente la información corológica de estas dos familias de arañas, relevante tanto a nivel de la región de estudio (Comunidad de Madrid) como a nivel Peninsular (ver Anexos y Jiménez-Valverde, 2002; Jiménez-Valverde *et al*., 2004; Jiménez-Valverde *et al*., 2006).

Figura 2.- A la izquierda, ejemplar de *Araneus diadematus* Clerck, 1758 (Araneidae); a la derecha, ejemplar de *Xysticus ferrigineus* Menge, 1876 (Thomisidae).

## APROXIMACIÓN AUTOECOLÓGICA: MODELOS PREDICTIVOS DE DISTRIBUCIÓN DE ESPECIES

Los modelos predictivos de distribución de especies han recibido especial atención en las dos últimas décadas. Las predicciones de ellos derivadas se han usado para explorar cuestiones a las que, de otra manera, hubiera sido prácticamente imposible aproximarse. Así, el uso más generalizado ha consistido en cuantificar y analizar la relación de la especie de interés con una serie de posible variables explicativas y emplear esa información para predecir los efectos que tendrán los cambios que en ellas se puedan producir (Gibson *et al*., 2004; Eyre & Buck, 2005; Jiménez, 2005; Sánchez-Cordero *et al*., 2005; Seoane *et al*., 2006). Los modelos predictivos de distribución también se han empleado para explorar hipótesis ecológicas y biogeográficas (Anderson *et al*., 2002; Lobo *et al*., 2006; Jiménez-Valverde *et al*., in press), hipótesis evolutivas (Peterson *et al*., 1999; Peterson & Holt, 2003; Wiens & Graham, 2005), para estudiar el efecto del cambio climático sobre las distribuciones (Peterson, 2003a; Thuiller *et al*., 2005), para predecir el rango geográfico de especies invasoras (Peterson & Vieglais, 2001; Peterson, 2003b) o para llevar a cabo selección de reservas (Araújo & Williams, 2000; Cabeza *et al*., 2004; Sánchez-Cordero *et al*., 2005), entro otros. Incluso con datos de distribución pobres y visiblemente sesgados, y empleando las técnicas más sencillas de modelización, los resultados de los modelos se han usado para resaltar las zonas carentes de información y sugerir así áreas donde enfocar futuros trabajos de campo (Jiménez-Valverde *et al*., 2006; Richardson *et al*., 2006). Los modelos predictivos han demostrado ser más fiables que los mapas de distribución publicados en guías de campo y atlas de distribución (Bustamante & Seoane, 2004). También han demostrado ser más fiables

que las hipótesis de distribución elaboradas por expertos (Seoane *et al.*, 2005; pero ver Pearce *et al.*, 2001). En definitiva, parece que nos encontramos ante una potente herramienta capaz de generar hipótesis comprobables y explorar patrones de distribución a partir de datos incompletos.

**La obsesión por la técnica. —** En los últimos años se ha desplegado un gran esfuerzo en el desarrollo de nuevas técnicas de modelización, con lo que actualmente la panoplia de métodos llega a ser abrumadora. Por una parte, existen estrategias para trabajar exclusivamente con datos de presencia: Bioclim (Busby, 1986, 1991), Domain (Carpenter *et al*., 1993), Analisis Factorial de Nicho (ENFA; Hirzel *et al.*, 2002), algoritmos genéticos (GARP; Stockwell & Peters, 1999), Fuzzy Bioclim (FEM; Robertson *et al*., 2004) o máxima entropía (MAXENT; Phillips *et al.*, 2006). Los métodos para trabajar con datos de presencia y ausencia también son muy diversos: Modelos Generalizados Lineales (GLM; McCullagh & Nelder, 1989), Modelos Generales Aditivos (GAM; Hastie & Tibshirani, 1990), Redes Neuronales (NNET; Fitzgerald & Lees, 1992), Árboles de Clasificación y Regresión (CART; Breiman *et al*., 1984) o técnicas de clasificación bayesiana (Termansen *et al.*, 2006), entre otros. Muchos estudios se han centrado en comparar el funcionamiento de las distintas técnicas (por ejemplo, Manel *et al*., 1999a, b; Elith & Burgman, 2002; Fertig & Reiners, 2002; Olden & Jackson, 2002; Thuiller *et al*., 2003; Brotons *et al*., 2004; Engler *et al*., 2004; Muñoz & Felicísimo, 2004; Segurado & Araújo, 2004; Elith *et al*., 2006), la mayoría llegando a la conclusión de que las técnicas más sofisticadas, aquellas que pueden establecer relaciones complejas entre las variable dependiente y las independientes, son las que ofrecen los mejores resultados. Araújo *et al*. (2005) y Pearson *et al.* (2006) llaman la atención sobre los efectos que tienen las numerosas fuentes de incertidumbre sobre los resultados de los modelos, mostrando la gran

variedad de patrones de cambio en los rangos geográficos frente al calentamiento global, muchos de ellos contrarios, que se pueden obtener para una misma especie empleando diferentes técnicas. Así, Araújo *et al*. (2005) proponen la realización de modelos de consenso, es decir, usar la media de las predicciones obtenidas con diferentes técnicas como la hipótesis de mayor poder predictivo.

Estos resultados no son del todo sorprendentes; por una parte, dada la inevitable falta de total independencia de los datos de validación con respecto a los de entrenamiento, es lógico que los métodos que tienden a sobreajustar muestren mejores resultados. Por otra parte, dado que la parametrización de las relaciones entre los factores y los datos de distribución varía con la técnica, y dado que unos mismos datos pueden ser modelizados mediante diferentes formulaciones, las predicciones de diferentes técnicas diferirán inevitablemente incluso cuando los resultados de las validaciones sobre datos independientes sean optimistas para todos los métodos (Van Nes & Scheffer, 2005). Es curioso, sin embargo, comprobar como, cuando se trabaja con buenos datos de distribución y con variables explicativas de las cuales se tiene algún indicio sobre su influencia en las especies de interés, las diferencias en la capacidad predictiva entre las técnicas se minimizan (Elith *et al*., 2006). Por tanto, parece que es fundamental trabajar con variables dependientes libres de errores y sesgos y con variables independientes que realmente ejerzan una influencia directa y causal sobre las especies. Sin embargo, se ha escrito poco sobre la influencia de estas fuentes de error en los resultados de los modelos. Además, si no se comprenden las causas que hacen que los modelos se comporten como lo hacen, las hipótesis generadas por éstos nunca podrán ser adecuadamente comprobadas (Van Nes & Scheffer, 2005). Da la sensación de que estamos cegados por la estadística, que ni

siquiera tratamos de comprender, y nos hemos olvidado de lo más importante, los datos biológicos.

**La variable dependiente. —** Si no vamos a elaborar un muestreo de campo, sino que vamos a trabajar con la información corológica ya existente, el primer paso es recopilar toda esa información sobre la distribución de la especies de interés. Para evitar errores en los datos corológicos hay que evaluar la credibilidad de las citas, algo especialmente importante si la especie es difícil de identificar y puede ser confundida con otros congéneres, situación relativamente frecuente en el caso de los artrópodos. Algunos autores han propuesto metodologías para la evaluación de los datos corológicos, por ejemplo, teniendo en cuenta los datos de presencia/ausencia de otras especies de los ensamblajes (Palmer *et al.*, 2003). Las citas dudosas deben ser atentamente consideradas antes de incluirlas en el proceso de modelización.

A pesar de que, como hemos visto anteriormente, existen técnicas capaces de modelizar usando únicamente datos de presencia, emplear ausencias restringe las predicciones allí donde es necesario, por lo que se ha argumentado que los modelos que emplean datos de ausencia son más precisos (Zaniewski *et al.* 2002; Engler *et al*. 2004). Sin embargo, usar eventos del tipo 1/0 (presencia/ausencia) implica tener en cuenta una serie de consideraciones.

En caso de no disponer de verdaderas ausencias, ha de evitarse la práctica habitual de incluir como ausencias todas aquellas localidades en las que hay ausencia de información (por ejemplo, Segurado & Araújo, 2004; Eyre *et al.*, 2005; Luoto *et al.*, 2005) ya que esto conllevará, inevitablemente, incluir falsas ausencias. Si no contamos con datos de ausencia procedentes de muestreos de campo, se puede recurrir a estrategias de generación de pseudo-ausencias, empleando para ello alguna técnica sencilla de modelización (Bioclim o ENFA, por ejemplo) para delimitar las áreas

ambientalmente alejadas del área de distribución conocida y así poder muestrear pseudo-ausencias con una alta probabilidad de ser verdaderas ausencias (Engler *et al*., 2004; Lobo *et al*., 2006). Sin embargo, ha de tenerse en cuenta que el concepto de lo que estamos modelizando variará dependiendo del tipo de ausencias usadas (Soberón & Peterson, 2005). En el primer caso, empleando verdaderas ausencias obtenidas con muestreos intensivos, nuestra predicción se aproximará a la distribución real de la especie. En el segundo caso, usando pseudo-ausencias alejadas del espacio ambiental de los puntos de presencia, nuestro modelo se aproximará más a la distribución potencial de la especie, es decir, a la distribución que podría tener en ausencia de las restricciones impuestas por factores históricos, geográficos o de otro tipo. Normalmente, las consideraciones acerca de la selección de los datos en función de los objetivos de la investigación no están claras en los trabajos de modelización predictiva, lo que inevitablemente genera errores de interpretación en los resultados.

Tanto el tamaño de muestra como la prevalencia (relación entre el número de presencias y el tamaño de muestra) de los datos son considerados como dos de los principales factores que afectan a la precisión de los modelos. Mientras que el tamaño de muestra ha recibido más atención, tanto por parte de los estadísticos (Freedman & Pee, 1989; Peduzzi *et al*., 1996; Steyerberg *et al*., 1999; Calvo & Domínguez, 2002) como por parte de los ecólogos (Pearce & Ferrier, 2000; Stockwell & Peterson, 2002; McPherson *et al*., 2004), el efecto de la prevalencia sobre los modelos no termina de estar claro y, generalmente, se asume que es mejor trabajar con prevalencias de 0.5 a fin de evitar sus supuestos efectos negativos en la parametrización de las funciones (Vaughan & Ormerod, 2003; McPherson *et al*., 2004).

Sin embargo, raramente se cuenta con el mismo número de presencias que de ausencias. Generalmente, si la especie es relativamente común, el número de

ausencias fiables suele ser escaso en comparación con los datos de presencia. Al contrario, cuando se trabaja con especies raras, el número de presencias puede llegar a ser muy bajo en comparación al número de ausencias disponibles. Esta desigualdad en los resultados del evento da lugar a un sesgo en las probabilidades medias de cada uno, estando sesgadas hacia el evento más común (Cramer, 1999). Esto implica que el punto de corte correcto empleado para convertir el mapa probabilístico producido en el proceso de modelización en uno de presencia/ausencia no sea 0.5, como intuitivamente parece (ver, por ejemplo, Manel *et al*., 1999b; Meggs *et al.*, 2004; Jiménez, 2005). Las implicaciones de este efecto matemático inevitable son importantes, ya que la transformación de las probabilidades en un mapa booleano es necesaria para computar los errores de comisión y omisión de los modelos. Curiosamente, únicamente conocemos un trabajo (Liu *et al*., 2005) que aborde específicamente una comparación de diferentes estrategias para calcular este punto de corte, empleando especies reales y redes neuronales como método de modelización. En el **capítulo 5** hemos estudiado el efecto de varios punto de corte en las predicciones de modelos regresivos (regresión logística) empleando una especie virtual, de tal manera que eliminamos cualquier otra fuente de error que añadiría incertidumbre a los resultados obtenidos.

Las probabilidades generadas por los modelos están sesgadas, por lo que no informan correctamente sobre la adecuación del hábitat (Rojas *et al.*, 2001) y han de ser reescaladas. De nuevo, únicamente existe un trabajo (Real *et al*., en prensa) en el que se aborde este aspecto de manera específica. Nosotros hemos discutido brevemente la cuestión en el **capítulo 6** y en el **capítulo 9** mostramos su aplicación.

Una vez considerado el inevitable efecto matemático de la prevalencia, ¿cuál es su verdadero efecto en los modelos predictivos?, ¿debemos remuestrear los datos,

reduciendo así el tamaño de muestra y desechando información, para trabajar con prevalencias de 0.5?, ¿podemos trabajar con especies raras? En el **capítulo 6** se aborda de manera teórica el efecto de la prevalencia y en el **capítulo 7** se estudia su efecto y su interacción con el tamaño de muestra empleando, de nuevo, una especie virtual y la regresión logística como método de modelización.

Mientras que una presencia suele ser un dato irrefutable (salvo errores de identificación ya comentados), una ausencia tiene siempre un grado de incertidumbre asociado. Una ausencia se puede deber a que la especie no ha sido buscada en la localidad de interés (falsas ausencias comunes en los atlas biológicos) o a que, a pesar de llevarse a cabo un muestreo, la especie no fue detectada. La distribución de estas falsas ausencias puede ser aletoria o mostrar una determinada estructura espacial. Teniendo en cuenta los sesgos presentes en nuestro conocimiento de la distribución biológica, es más que probable que la última situación sea el caso habitual. Es llamativo comprobar como, a pesar de que la existencia de falsas ausencias es una fuente de error omnipresente en cualquier proceso de modelización, su efecto no ha sido prácticamente estudiado. Tyre *et al*. (2003) y Gu & Swihart (2004) demuestran como la presencia de falsas ausencias provoca errores en las estimas de los parámetros de las funciones. Gu & Swihart (2004) muestran como, en regresiones logísticas, si las falsas ausencias están asociadas a ciertos valores de alguna de las variables predictoras, entonces la importancia de estas variables será sobreestimada. En el **capítulo 8** estudiamos el efecto de las falsas ausencias, con y sin estructura espacial, en las predicciones de los modelos.

**Las variables independientes. —** Una vez que se dispone de la variable dependiente libre de errores, hay que analizar las variables independientes o explicativas. Lo ideal es trabajar con variables que se sepa ejercen una acción directa

causal en la delimitación del rango geográfico de la especie. Sin embargo, tanto la falta de estudios fisiológicos como la disponibilidad de variables almacenadas en capas temáticas georreferenciadas, limitan el rango de las variables a utilizar.

La mayoría de las variables ambientales muestran un elevado grado de correlación entre sí, lo se supone un problema para los análisis, especialmente, aquellos que se basan en técnicas de regresión. La inclusión de dos variables altamente colineales puede implicar estimas imprecisas, tanto de sus parámetros como de los del resto de variables (Bagley *et al*., 2001). Por tanto, un análisis previo de correlación nos ayudará a identificar grupos de factores redundantes. Por otra parte, si contamos con un elevado número de variables candidatas a ser incluidas en el modelo, resultará interesante hacer una reducción previa. La inclusión de un alto número de factores en los modelos puede provocar problemas en las estimas de los parámetros y de sobreajuste, reduciendo la capacidad de extrapolación de los modelos (Harrel *et al.*, 1996; Reineking & Schröder, 2003; Fig. 3). Para reducir el número de variables se pueden seguir varios criterios; el primero consiste en que las variables candidatas a

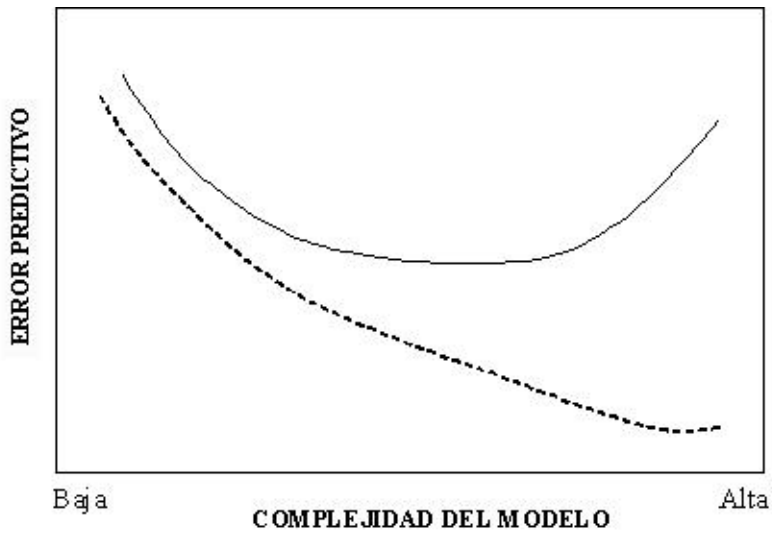



Figura 3.- Relación entre la complejidad del modelos y el error predictivo en los datos de entrenamiento (línea de rayas) y en los datos de validación independientes (línea continua) (modificado de Reineking & Schröder, 2003).

eliminarse deben mostrar un alto grado de correlación con otras, de manera que se elimine la información redundante. Para decidirse por unas o por otras, y no dejar la elección al azar, se pueden realizar analisis univariantes entre los factores y los datos de presencia/ausencia con el fin de identificar las variables menos explicativas y/o las que muestran relaciones poco realistas desde un punto de vista biológico.

**Los ejemplos: distribución potencial de *Macrothele calpeiana* (Walckenaer, 1805) en la Península Ibérica. —** Todos estos aspectos (y otros tratados en el apartado anterior referente a la parte sinecológica de la tesis) han sido considerados en los **capítulos 9 y 10**, en los que se ha modelizado la distribución potencial de la araña *Macrothele calpeiana* (Walckenaer, 1805) (Fig. 4). Esta especie pertenece a la familia Hexathelidae, es un endemismo Ibérico y se encuentra incluida en el Convenio de Berna y en la Directiva Hábitats (Ferrández, 2004). Es una especie que cuenta con 6 núcleos de población supuestamente aislados en el sur de la Península Ibérica, habita en nidos de seda con forma de embudo que sitúa principalmente bajo piedras (Fig. 5) y se le supone una baja capacidad de dispersión. En el **capítulo 9** se ha afrontado la predicción de la distribución potencial en la Península Ibérica, empleando para ello variables exclusivamente climáticas, con el fin de elaborar una hipótesis sobre la distribución de especie. El modelo obtenido se ha extrapolado al Norte de África y al resto de la región Mediterránea con el fin de identificar áreas que posean condiciones ambientales idóneas para la araña. Además, se ha evaluado el posible efecto del cambio climático sobre la distribución potencial en la Península y en el Norte de África. En el **capítulo 10** se ha elaborado un modelo predictivo a una menor escala (resolución y extensión) con el objetivo de estimar la importancia relativa de distintos grupos de factores en la determinación del rango de

distribución de *M. calpeiana* (climáticos, vigor vegetal y usos del suelo) empleando para ello, además, técnicas de partición y jerarquización de la variabilidad explicada.

Figura 4.- Ejemplar de *Macrothele calpeiana* (Walckenaer, 1805).

Figura 5.- Nido de *Macrothele calpeiana* (Walckenaer, 1805). Se observa, en primer plano, la parte aérea de la tela. Al fondo, el túnel de seda que, en este caso, se escondía debajo de una piedra.

## OBJETIVOS DE LA TESIS DOCTORAL

Los objetivos concretos que persigue la presente tesis doctoral son:

I. Definir un protocolo de muestreo para conseguir inventarios fiables de las familias de arañas Araneidae y Thomisidae en parcelas de 1 $km^2$.

II. Elaborar una hipótesis sobre los factores más influyentes a la hora de determinar el patrón de riqueza de las dos familias en la Comunidad de Madrid.

III. Estudiar el efecto i) del punto de corte para convertir mapas continuos de probabilidad en mapas booleanos, ii) de la prevalencia y iii) de las falsas ausencias en los modelos de distribución potencial de especies.

IV. Desarrollar un modelo de distribución potencial de *Macrothele calpeiana* (Walckenaer, 1805) en la Península Ibérica con el fin elaborar una hipótesis sobre los factores determinantes de su rango geográfico y de ilustrar minuciosamente los pasos necesarios para elaborar un modelo individual.

## ESTRUCTURA DE LA TESIS DOCTORAL

La presente tesis doctoral consta de dos partes claramente diferenciadas:

I. La primera parte aborda un análisis sinecológico, en el cual se desarrolla un protocolo de muestreo para las familias de arañas Araneidae y Thomisidae, se propone un método de selección de puntos de muestreo y, finalmente, se elabora un modelo predictivo de riqueza de estas dos familias en la Comunidad de

Madrid con el fin de realizar una primera aproximación a los posibles determinantes del patrón de variación en la diversidad biológica de estos grupos.

II. La segunda parte presenta un estudio autoecológico en el que, primeramente, se abordan tres cuestiones metodológicas que hemos considerado especialmente relevantes: la determinación del punto de corte para convertir un mapa de probabilidades predictivo de distribución en uno booleano de presencia/ausencia, la influencia sobre la capacidad predictiva de los modelos del empleo de una variables dependiente (presencia/ausencia) con los eventos no equilibrados (prevalencias sesgadas), y la influencia de las falsas ausencias y de su estructura espacial. Finalmente, se elaboran dos modelos de distribución potencial, a dos escalas (resolución y extensión) espaciales diferentes, de una especie de araña con interés conservacionista, *Macrothele calpeiana*, endémica de la Península Ibérica.

**Primera Parte: Riqueza de las familias Araneidae y Thomisidae en la Comunidad de Madrid. —** Esta primera parte consta de cuatro capítulos:

I. *Un protocolo de muestreo combinado para la estimación de los ensamblajes de Araneidae y Thomisidae (Aracnida, Araneae).*

En este capítulo se comparan diferentes técnicas de muestreo, ampliamente utilizadas para la captura de arañas, con el fin de definir una combinación que permita obtener una representación lo mas fiel posible de la riqueza de especies de las familias Araneidae y Thomisidae en parcelas de 1 $km^2$. Se estudia también el esfuerzo de muestreo necesario para obtener tal inventario.

II. *Definiendo protocolos de muestreo óptimos de arañas (Araneae, Araneidae y Thomisidae): estimación de la riqueza específica, cobertura estacional y contribución de los individuos juveniles a la riqueza y composición de especies.*

Una vez establecida la combinación ideal de métodos de muestreo y el esfuerzo de muestreo necesario, en este capítulo se aborda la capacidad de diferentes diseños temporales de muestreo para obtener estimas fiables sobre la diversidad de arañas en una parcela de 1 $km^2$. Además, se estudia el efecto de la inclusión o no de individuos juveniles en los análisis sobre las estimas.

III. *Un método sencillo para seleccionar puntos de muestreo con el objeto de inventariar taxones hiperdiversos: el caso práctico de las familias Araneidae y Thomisidae (Araneae) en la Comunidad de Madrid, España.*

En este capítulo se presenta un método para la selección de puntos de muestreo para estudiar la riqueza de especies de un territorio determinado. Haciendo una estima del esfuerzo de muestreo que es posible invertir durante el desarrollo de la presente tesis doctoral, el método se aplica a la selección de los puntos de muestreo para el estudio de la riqueza de las familias Araneidae y Thomisidae en la Comunidad de Madrid.

IV. *Determinantes de la riqueza local de arañas (Araneidae y Thomisidae) en una escala regional: clima y altitud vs. estructura de hábitat.*

En este capítulo se estudian los factores determinantes de la riqueza local de Araneidae y Thomisidae en la Comunidad de Madrid. Tras elaborar los inventarios de 15 localidades de 1 $km^2$ en la región, y empleando una serie de variables descriptoras del medio, se modeliza la riqueza especifica

empleando Modelos de Regresión y se comparan los efectos relativos de tres grupos de variables empleando la partición de la varianza.

**Segunda Parte: Distribución potencial del endemismo ibérico *Macrothele calpeiana* (Walckenaer, 1805) (Araneae, Hexathelidae). —** Esta segunda parte de la tesis consta de 6 capítulos:

*I.* *Criterios para seleccionar el punto de corte con el fin de convertir mapas continuos de probabilidad de presencia a mapas booleanos de presencia/ausencia.*

En este capítulo se comparan diferentes criterios para fijar un punto de corte en los modelos predictivos continuos, derivados de una regresión logística, para convertirlos a mapas de presencia/ausencia. Empleamos una especie virtual con el fin de controlar todas las posible fuentes de error (tamaño muestral, falsos datos de distribución, variables espurias, etc.).

*II.* *El fantasma de los eventos no equilibrados en los modelos predictivos de distribución de especies*.

Aquí plasmamos algunas ideas teóricas sobre el efecto que sobre la capacidad predictiva de los modelos tiene el uso de muestras no equilibradas, es decir, muestras con prevalencias sesgadas en la variable dependiente.

*III.* *Efectos de la prevalencia y de su interacción con el tamaño de muestra en los modelos de distribución de especies: necesitamos muchos más datos de ausencia.*

En este capítulo estudiamos el efecto que la prevalencia tiene sobre los modelos predictivos, nuevamente empleando una especie virtual con el fin

de controlar todas las posibles fuentes de error y aislar, de esta manera, el verdadero efecto de la prevalencia.

*IV.* *El efecto de las falsas ausencias en los modelos predictivos de distribución.*

Es este capítulo se estudia el efecto de las falsas ausencias, con y sin estructura espacial, en la capacidad predictiva de los modelos. Volvemos a estudiar estos efectos con una especie virtual para aislar los verdaderos efectos de la fuente de error que deseamos testar.

*V.* *Distribución potencial de la araña* Macrothele calpeiana *(Walckenaer, 1805) (Araneae, Hexathelidae) en la Península Ibérica, extraplación al Norte de África y a la región Mediterránea, y evaluación del impacto del cambio climático,* y

*VI.* *Factores determinantes de la distribución del endemismo ibérico* Macrothele calpeiana *(Walckenaer, 1805) (Araneae, Hexathelidae).*

Estos dos últimos capítulos muestran el proceso completo de modelización de una especie, teniendo en cuanta los aspectos específicamente analizados en la tesis doctoral y otros que han sido considerados de especial relevancia y tratados en esta Introducción.

## REFERENCIAS BIBLIOGRÁFICAS

Adler, P. B. & Lauenroth, W. K. (2003) The power of time: spatiotemporal scaling of species diversity. *Ecology Letters*, **6,** 749-756.

Anderson, R. P., Gómez-Laverde, M. & Peterson, A. T. (2002) Geographical distributions of spiny pocket mice in South America: insights from predictive models. *Global Ecology and Biogeography*, **11**, 131-141.

Andriamampianina, L., Kremen, C., Vane-Wright, D., Lees, D., Razafimahatratra, V. (2000) Taxic richness patterns and conservation evaluation of Madagascan tiger beetles (Coleoptera: Cicindelidae). *Journal of Insect Conservation*, **4**, 109-128.

Araújo, M. B. (1999) Distribution patterns of biodiversity and the design of a representative reserve network in Portugal. *Diversity and Distributions*, **5**, 151-163.

Araújo, M. B. & Williams, P. H. (2000) Selecting areas for species persistence using occurrence data. *Biological Conservation*, **96**, 331-345.

Araújo, M. B., Densham, P. J., Williams, P. H. (2004) Representing species in reserves from patterns of assemblage diversity. *Journal of Biogeography*, **31**, 1037-1050.

Araújo, M. B., Whittaker, R. J., Ladle, R. J. & Erhard, M. (2005) Reducing uncertainty in projections of extinction risk from climate change. *Global Ecology and Biogeography*, **14**, 529-538.

Bagley, S. C., White, H. & Golomb, B. A. (2001) Logistic regression in the medical literature: standards for use and reporting, with particular attention to one medical domain. *Journal of Clinical Epidemiology*, **54**, 979-985.

Baselga, A., Hortal, J., Jiménez-Valverde, A., Gómez, J. F. & Lobo, J. M. Which leaf beetles have not yet been described? Determinants of the description of Western Palaeartic *Aphtona* species (Coleoptera: Chrysomelidae). *Biodiversity and Conservation*, en prensa.

Breiman, L., Friedman, J. H., Olshen, R. A. & Stone, C. J. (1984) *Classification and regression trees*. Wadsworth International Group, Belmont, CA.

Brotons, L., Thuiller, W., Araújo, M. B. & Hirzel, A. H. (2004) Presence-absence versus presence-only modelling methods for predicting bird habitat suitability. *Ecography*, **27**, 437-448.

Bunge, J. & Fitzpatrick, M. (1993) Estimating the number of species: a review. *Journal of the American Statistical Association*, **88**, 364-373.

Busby, J. R. (1986) A biogeographical analysis of *Nothofagus cunninghamii* (Hook.) Oerst. in southern Australia. *Austalian Journal of Ecology*, **11**, 1-7.

Busby, J. R. (1991) BIOCLIM – A Bioclimate Analysis and Prediction System. En *Nature Conservation: Cost Effective Biological Surveys and Data* Analysis, eds. C. R. Margules & M. P. Austin, pp. 64-68. CSIRO, Australia.

Bustamante, J. & Seoane, J. (2004) Predicting the distribution of four species of raptors (Aves: Accipitridae) in southern Spain: statistical models work better than existing maps. *Journal of Biogeography*, **31**, 295-306.

Cabeza, M., Araújo, M. B., Wilson, R. J., Thomas, C. D., Cowley, M. J. R. & Moilanen, A. (2004) Combining probabilities of occurrence with spatial reserve design. *Journal of Applied Ecology*, **41**, 252-262.

Cabrero-Sañudo, F. J. & Lobo, J. M. (2003) Estimating the number of species not yet described and their characteristics: the case of western Palaearctic dung beetle species (Coleoptera, Scarabaeoidea). *Biodiversity and Conservation*, **12**, 147-166.

Calvo, M. O. & Domínguez, A. C. (2002) Regresión logística no condicionada y tamaño de muestra: una revisión bibliográfica. *Revista Española de Salud Pública*, **76**, 85-93.

Carpenter, G., Gillison, A. N. & Winter, J. (1993) DOMAIN: a flexible modeling procedure for mapping potential distributions of plants and animals. *Biodiversity and Conservation*, **2**, 667-680.

Chapin, F. S., Zavaleta, E. S., Eviner, V. T., Naylor, R. L., Vitousek, P. M., Reynolds, H. L., Hooper, D. U., Lavorel, S., Sala, O. E., Hobbie, S. E., Mack, M. V. & Díaz, S. (2000) Consequences of changing biodiversity. *Nature*, **405**, 234-242.

Chatfield, C. (1995) Model uncertainty, data mining and statistical inference. *Journal of the Royal Statistical Society*, **158**, 419-466.

Colwell, R. K. & Coddington, J. A. (1994) Estimating terrestrial biodiversity through extrapolation. *Philosophical Transactions of the Royal Society, series* B, **345**, 101-118.

Cramer, J. S. (1999) Predictive performance of binary logit model in unbalanced samples. *Journal of the Royal Statistical Society: Series D (The Statistician)*, **48**, 85-94.

Diamond, J. M. (1985) How many unknown species are yet to be discovered? *Nature*, **315**, 538-539.

Dennis, R. L. H. (2001) Progressive bias in species status is symptomatic of fine-grained mapping units subject to repeated sampling. *Biodiversity and Conservation*, **10**, 483-494.

Dennis, R. L. H., Sparks, T. H. & Hardy, P. B. (1999) Bias in butterfly distribution maps: the effects of sampling effort. *Journal of Insect Conservation*, **3**, 33-42.

Dennis, R. L. H. & Thomas, C. D. (2000) Bias in butterfly distribution maps: the influence of hot spots and recorder's home range. *Journal of Insect Conservation*, **4**, 73-77.

Dobson, A. P., Rodríguez, J. P., Roberts, W. M. & Wilcove, D. S. (1997) Geographic distribution of endangered species in the United States. *Science*, **275**, 550-553.

Dunn, R. R. (2005) Modern insect extinctions, the neglected majority. *Conservation Biology*, **19**, 1030-1036.

Elith, J. & Burgman, M. (2002) Predictions and their validations: rare plants in the Central Highlands, Victoria, Australia. En *Predicting species occurrences: Issues of accuracy and scale*, ed. J. M. Scott, M. L. Morrison & P. J. Heglund, pp. 303-313. Island Press, Covelo, CA.

Elith, J., Graham, C. H., Anderson, R. P., Dudík, M., Ferrier, S., Guisan, A., Hijmans, R. J., Huettmann, F., Leathwick, J. R., Lehmann, A., Li, J., Lohmann, L.G., Loiselle, B. A., Manion, G., Moritz, C., Nakamura, M., Nakazawa, Y., Overton, J. M., Peterson, A. T., Phillips, S. J., Richardson, K., Scachetti-Pereira, R., Schapire, R. E., Soberón, J., Williams, S., Wisz, M. S. & Zimmermann, N. E. (2006) Novel methods improve prediction of species' distributions from occurrence data. *Ecography*, **29**, 129-151.

Engler, R., Guisan, A & Rechsteiner, L. (2004) An improved approach for predicting the distribution of rare and endangered species from occurrence and pseudo-absence data. *Journal of Applied Ecology*, **41**, 263-274.

Eyre, T. J. & Buck, R. G. (2005) The regional distribution of large gliding possums in southern Queensland, Australia. I. The yellow-bellied glider (*Petaurus australis*). *Biological Conservation*, **125**, 65-86.

Eyre, M. D., Rushton, S. P., Luff, M. L. & Telfer, M. G. (2005) Investigating the relationship between the distribution of British ground beetle species (Coleoptera, Carabidae) and temperature, precipitation and altitude. *Journal of Biogeography*, **32**, 973-983.

Ferrández, M. A. (2004) *Macrothele calpeiana* (Walckenaer, 1805). Situación actual y perspectivas. *Munibe (Suplemento)*, **21**, 154-161.

Ferrier, S. (2002) Mapping spatial pattern in biodiversity for regional conservation planning: Where to from here? *Systematic Biology*, **51**, 331-363.

Ferrier, S., Drielsma, M., Manion, G., Watson, G. (2002a) Extended statistical approaches to modelling spatial pattern in biodiversity in northeast New South Wales. II. Community-level modelling. *Biodiversity and Conservation*, **11**, 2309-2338.

Ferrier, S., Watson, G., Pearce, J., Drielsma, M. (2002b) Extended statistical approaches to modelling spatial pattern in biodiversity in northeast New South Wales. I. Species-level modelling. *Biodiversity and Conservation*, **11**, 2275-2307.

Fertig, W. & Reiners, W. A. (2002) Predicting presence/absence of plant species for range mapping: a case study from Wyoming. En *Predicting species occurrences: Issues of accuracy and scale*, ed. J. M. Scott, M. L. Morrison & P. J. Heglund, pp. 483-489. Island Press, Covelo, CA.

Fielding, A. H., & Bell, J. F. (1997) A review of methods for the assessment of prediction errors in conservation presence/absence models. *Environmental Conservation*, **24**, 38-49.

Fitzgerald, R. W. & Lees, B. G. (1992) The application of neural networks to the floristic classification of remote sensing and GIS data in complex terrain. En *Proceedings of the XVII Congress ASPRS*, ed. American Society of Photogrametry and Remote Sensing, pp. 570-573. Bethesda, MD.

Freedman L. S. & Pee, D. (1989) Return to a note on screening regression equations. *American Statistician*, **43,** 279-282.

García-Barros, E., García-Pereira, P. & Munguira, M. L. (2000) The geographic distribution and state of butterfly faunistic studies in Iberia (Lepidoptera Papilionoidea Hesperioidea). *Belgian Journal of Entomology*, **2**, 111-124.

Gaston, K. J. (1996) Species richness: measure and measurement. En *Biodiversity. A biology of numbers and difference*, ed. K. J. Gaston, pp. 77-113. Blackwell Science, Oxfrod.

Gaston, K. J. & May, R. M. (1992) The taxonomy of taxonomists. *Nature*, **356**, 281-282.

Gibson, L. A., Wilson, B. A., Cahill, D. M. & Hill, J. (2004) Spatial prediction of rufous bristlebird habitat in a coastal heathland: a GIS-based approach. *Journal of Applied Ecology*, **41**, 213-223.

Gladstone, W. (2002) The potential value of indicator groups in the selection of marine reserves. *Biological Conservation*, **104**, 211-220.

Gotelli, N. J. & Colwell, R. K. (2001) Quantifying biodiversity: procedures and pitfalls in the measurement and comparison of species richness. *Ecology Letters*, **4**, 379-391.

Graham, C. H., Ferrier, S., Huettman, F., Moritz, C. & Peterson, A.T. (2004) New developments in museum-based informatics and applications in biodiversity analysis. *Trends in Ecology and Evolution*, **19**, 497-503.

Groombridge, B. & Jenkins, M. D. (2002) *World Atlas of Biodiversity. Earth´s living resources in the 21$_{st}$ century*. University of California Press, California.

Gu, W. & Swihart, R. K. (2004) Absent or undetected? Effects of non-detection of species occurrence on wildlife–habitat models. *Biological Conservation*, **116**, 195–203.

Guisan, A. & Zimmermann, N. E. (2000) Predictive habitat distribution models in ecology. *Ecological Modelling*, **135**, 147-186.

Guisan, A. & Thuiller, W. (2005) Predicting species distributions: offering more than simple habitat models. *Ecology Letters*, **8**, 993-1009.

Hammond, P. M. (1994) Practical approaches to the estimation of the extent of biodiversity in speciose groups. *Philosophical Transactions of the Royal Society, series B*, **345**, 119-136.

Harper, J. L. & Hawksworth, D. L. (1995). Preface. En *Biodiversity. Measurement and estimation*, ed. D. L. Hawksworth, pp. 5-12. Chapman & Hall, Oxford.

Harrel, F. E., Lee, K. L. & Mark, D. B. (1996) Multivariable prognostic models: issues in developing models, evaluating assumptions and adequacy, and measuring and reducing errors. *Statistics in Medicine*, **15**, 361-387.

Hastie, T. J. & Tibshirani, R. J. (1990) *Generalized Additive Models*. Chapman & Hall, London.

Hirzel, A. H., Hausser, J., Chessel, D. & Perrin, N. (2002) Ecological niche factor analysis: how to compute habitat suitability maps without absence data? *Ecology*, **83**, 2027-2036.

Hortal, J. & Lobo, J. M. (2001) A preliminary methodological approach to model the spatial distribution of biodiversity attributes. En *Spatio-temporal modeling of environmental processes: proceedings of the 1st Spanish Workshop of Spatio-temporal Modelling of Environmenatl Processes*, eds. J. Mateu & F. Montes, pp. 211-239. Plubicacions de la Universitat Jaume I, Collecció "Treballs d´Informàtica I Tecnologia" 10, Castelló de la Plana.

Hortal, J. & Lobo, J. M. (2002) Una metodología para predecir la distribución espacial de la diversidad biológica. *Ecología (n.s.),* **16,** 151-178 + 14 figuras.

Hortal, J. & Lobo, J. M. (2005) An ED-based protocol for optimal sampling of biodiversity. *Biodiversity and Conservation*, **14**, 2913-2947.

Hortal, J., Lobo, J. M. & Martín-Piera, F. (2001) Forecasting insect species richness scores in poorly surveyed territories: the case of the Portuguese dung beetles (Col. Scarabaeinae). *Biodiversity and Conservation*, **10**, 1343-1367.

Hortal, J., Lobo, J. M., Martín-Piera, F. (2003) Una estrategia para obtener regionalizaciones bióticas fiables a partir de datos incompletos: el caso de los Escarabeidos (Coleoptera) Ibérico-Baleares. *Graellsia*, **59**, 331-344.

Hortal, J., Garcia-Pereira, P., García-Barros, E. (2004) Butterfly species richness in mainland Portugal: Predictive models of geographic distribution patterns. *Ecography*, **27**, 68-82.

Howard, P. C., Viskanic, P., Davenport, T. R. B., Kigenyi, F. W., Baltzer, M., Dickinson, C. J., Lwanga, J. S., Matthews, R. A. & Balmford, A. (1998) Complementarity and the use of indicator groups for reserve selection in Uganda. *Nature*, **394**, 472-475.

Huston, M. A. (1994) *Biological Diversity. The coexistence of species on changing landscapes*. Cambridge University Press, Cambridge.

Johnston, C. A. (1998) *Geographic Information Systems in Ecology*. Blackwell Science, Oxford.

Jiménez, I. (2005) Development of predictive models to explain the distribution of the West Indian manatee *Trichechus manatus* in tropical watercourses. *Biological Conservation*, **125**, 491-503.

Jiménez-Valverde, A. (2002) Presencia en la Comunidad de Madrid (España central) del endemismo ibérico *Ozyptila umbraculorum* Simon, 1832 (Araneae, Thomisidae). *Revista Ibérica de Aracnología*, **6**, 225-227.

Jiménez-Valverde, A. (2005) Contribución al conocimiento de los saltícidos (Araneae, Salticidae) de la Comunidad de Madrid (España central). *Boletín de la Sociedad Entomológica Aragonesa,* **37**, 289-296.

Jiménez-Valverde, A. & Hortal, J. (2003) Las curvas de acumulación de especies y la necesidad de evaluar la calidad de los inventarios biológicos. *Revista Ibérica de Aracnología*, **8**, 151-161.

Jiménez-Valverde, A., Barriga Bernal, J. C. & Morano, E. (2004) Datos interesantes sobre la distribución de *Araniella opisthographa* (Kulczynski, 1905) y *A. inconspicua* (Simon, 1874) (Araneae: Araneidae) en la Península Ibérica. *Revista Ibérica de Aracnología*, **9**, 269-270.

Jiménez-Valverde, A. & Ortuño, V. M. The history of endemic Iberian ground beetle description (Insecta, Coleoptera, Carabidae): which species were described first? *Acta Oecologica*, en prensa.

Jiménez-Valverde, A., Lobo, J. M. & López Martos, M. L. (2006) Listado de especies actualizado de los araneidos y tomisidos (Araneae, Araneidae & Thomisidae) de la Comunidad de Madrid: mapas de distribución conocida, potencial y patrones de riqueza. *Graellsia*, en prensa.

Jiménez-Valverde, A., Ortuño, V. M. & Lobo, J. M. Exploring the distribution of *Sterocorax* Ortuño, 1990 (Coleoptera, Carabidae) species in the Iberian Peninsula. *Journal of Biogeography*, en prensa.

Lawton, J. H. & May, R. M. (1995) *Extinction Rates*. Oxford University Press, Oxfrod.

Legendre, P. & Legendre, L. (1998) *Numerical Ecology.* Elsevier, Amsterdam.

Liu, C., Berry, P. M., Dawson, T. P. & Pearson, R. G. (2005) Selecting thresholds of occurrence in the prediction of species distributions. *Ecography*, **28,** 385-393.

Lobo, J. M. (2000) ¿Es posible predecir la distribución geográfica de las especies basándonos en variables ambientales? En *Hacia un Proyecto CITED para el Inventario y Estimación de la Diversidad Entomológica en Iberoamérica: PrIBES-2000*, eds. F. Martín-Piera, J. J. Morrone & A. Melic, pp. 55-68. SEA, Zaragoza.

Lobo, J. M. & Martín-Piera, F. (2002) Searching for a predictive model for species richness of Iberian dung beetle based on spatial and environmental variables. *Conservation Biology*, **16**, 158-173.

Lobo, J. M., Castro, I. & Moreno, J. C. (2001) Spatial and environmental determinants of vascular plant species richness distribution in the Iberian Peninsula and Balearic Islands. *Biological Journal of the Linnean Society*, **73**, 233-253.

Lobo, J. M., Jay-Robert, P. & Lumaret, J.-P. (2004) Modelling the species richness distribution for French Aphodiidae (Coleoptera, Scarabaeoidea). *Ecography*, **27**, 145-156.

Lobo, J. M., Verdú, J. R. & Numa, C. (2006). Environmental and geographical factors affecting the Iberian distribution of flightless *Jekelius* species (Coleoptera: Geotrupidae). *Diversity and Distributions*, **12**, 179-188.

Lomborg, B. (2001) *The Skeptical Environmentalist. Measuring the Real State of the World*. Cambridge University Press, Cambridge.

Luoto, M., Pöyry, J., Heikkinen, R. K. & Saarinen, K. (2005). Uncertainty of bioclimate envelope models based on the geographical distribution of species. *Global Ecology and Biogeography*, **14**, 575-584.

Manel, S., Dias, J.-M. & Ormerod, S. J. (1999a) Comparing discriminant analysis, neural networks and logistic regression for predicting species distributions: a case study with a Himalayan river bird. *Ecological Modelling*, **120**, 337-347.

Manel, S., Dias, J.-M., Buckton, S. T. & Ormerod, S. J. (1999b) Alternative methods for predicting species distribution: an illustration with Himalayan river birds. *Journal of Applied Ecology*, **36**, 734-747.

Margules, C. R., Nicholls, A. O. & Pressey, R. L. (1988) Selecting networks of reserves to maximise biological diversity. *Biological Conservation*, **43**, 63-76.

Martín-Piera, F. (2001) Area networks for conserving Iberian insects: A case study of dung beetles (col., Scarabaeoidea). *Journal of Insect Conservation*, **5**, 233-252.

May, R. M. (1988) How many species are there on Earth? *Science*, **241**, 1441-1449.

May, R. M. (1994) Conceptual aspects of the quantification of the extent of biological diversity. *Philosophical Transactions of the Royal Society, series B*, **345**, 13-20.

Mac Nally, R. (2000) Regression and model-building in conservation biology, biogeography and ecology: the distinction between – and reconciliation of – "predictive" and "explanatory" models. *Biodiversity and Conservation*, **9**, 655–671.

Mac Nally, R. (2002) Multiple regression and inference in ecology and conservation biology: further comments on retention of independent variables. *Biodiversity and Conservation*, **11**, 1397-1401.

McCullagh, P. & Nelder, J. A. (1989) *Generalized linear models*. Chapman & Hall, London.

McPherson, J. M., Jetz, W. & Rogers, D. J. (2004) The effects of species´ range sizes on the accuracy of distribution models: ecological phenomenon or statistical artefact? *Journal of Applied Ecology*, **41**, 811-823.

Medellín, R. A. & Soberón, J. (1999) Predictions of mammal diversity on tour land masses. *Conservation Biology*, **13**, 143-149.

Meggs, J. M., Munks, S. A., Corkrey, R. & Richards, K. (2004) Development and evaluation of predictive habitat models to assist the conservation planning of a threatened lucanid beetle, *Hoplogonus simsoni*, in north-east Tasmania. *Biological Conservation*, **118**, 501-511.

Melic, A. (2001) Arañas endémicas de la Península Ibérica e Islas Baleares (Arachnida: Araneae). *Revista Ibérica de Aracnología*, **4**, 35-92.

Miller, R. I. (1994) *Mapping the diversity of nature*. Chapman & Hall, London.

Morano, E. (2004) Introducción a la diversidad de las arañas Iberobaleares. *Munibe (suplemento)*, **21**, 92-137.

Muñoz, J. & Felicísimo, Á. M. (2004) Comparison of statistical methods commonly used in predictive modelling. *Journal of Vegetation Science*, **15**, 285-292.

Nicholls, A. O. (1989) How to make biological surveys go further with Generalised Linear Models. *Biological Conservation*, **50**, 51-75.

Nielsen, S. E., Johnson, C. J., Heard, D. C. & Boyce, M. S. (2005) Can models of presence-absence be used to scale abundance? Two case studies considering extreme in life history. *Ecography*, **28**, 197-208.

Olden, J. D. & Jackson, D. A. (2000) Torturing data for the sake of generality: how valid are our regression models? *Ecoscience*, **7**, 501-510.

Olden, J. D. & Jackson, D. A. (2002) A comparison of statistical approaches for modeling fish species distributions. *Freshwater Biology*, **47**, 1976-1995.

Olden, J. D., Jackson, D. A. & Peres-Neto, P. R. (2002) Predictive models of fish species distributions: a note on proper validation and chance predictions. *Transactions of the American Fisheries Society*, **131**, 329-336.

Palmer, M. Gómez-Pujol, L., Pons, G. X., Mateu, J. & Linde, M. (2003) Noisy data and distribution maps: the example of *Phylan semicostatus* Mulsant and Rey, 1854 (Coleoptera, Tenebrionidae) from Serra de Tramontana (Mallorca, Western Mediterranean). *Graellsia*, **59**, 389-398.

Patterson, B. D. (1994) Accumulating knowledge on the dimensions of biodiversity: systematic perspectives on Neotropocal mammals. *Biodiversity Letters*, **2**, 79-86.

Patterson, B. D. (2000) Patterns and trends in the discovery of new Neotropical mammals. *Diversity and Distributions*, **6**, 145-151.

Pearce, J. & Ferrier, S. (2000) Evaluating the predictive performance of habitat models developed using logistic regression. *Ecological Modelling*, **133**, 225-245.

Pearce, J. & Ferrier, S. (2001) The practical value of modeling relative abundance of species for regional conservation planning: a case study. *Biological Conservation*, **98**, 33-43.

Pearce, J. L., Cherry, K., Drielsma, M., Ferrier, S. & Whish, G. (2001) Incorporating expert opinion and fine-scale vegetation mapping into statistical models of faunal distribution. *Journal of Applied Ecology*, **38**, 412-424.

Pearson, R. G., Thuiller, W., Araújo, M. B., Martínez-Meyer, E., Brotons, L., McClean, C., Miles, L., Segurado, P., Dawson, T. P. & Lees, D. C. (2006) Model-based uncertainty in species range prediction. *Journal of Biogeography*, en prensa.

Peduzzi. P., Concato, J., Kemper, E., Holford, T. R. & Feinstein, A. R. (1996) A simulation study of the number of events per variable in logistic regression analysis. *Journal of Clinical Epidemiology*, **49**, 1373-1379.

Peterson, A. T. (2003a) Projected climate change effects on Rocky Mountain and Great Plains birds: generalities of biodiversity consequences. *Global Change Biology*, **9**, 647-655.

Peterson, A. T. (2003b) Predicting the geography of species´ invasions via ecological niche modelling. *The Quarterly Review of Biology*, **78**, 419-433.

Peterson, A. T. & Holt, R. D. (2003) Niche differentiation in Mexican birds: using point occurrences to detect ecological innovation. *Ecology Letters*, **6**, 774-782.

Peterson, A. T. & Vieglais, D. A. (2001) Predicting species invasions using ecological niche modeling: new approaches from bioinformatics attack a pressing problem. *BioScience*, **51**, 363-371.

Peterson, A. T., Soberón, J. & Sánchez-Cordero, V. (1999) Conservatism of ecological niches in evolutionary time. *Science*, **285**, 1265-1267.

Phillips, S. J., Anderson, R. P. & Schapire, R. E. (2006) Maximum entropy modeling of species geographic distributions. *Ecological Modelling*, **190**, 231-259.

Pimm, S. L. & Raven, P. (2000) Extinction by numbers. *Nature*, **403**, 843-845.

Pimm, S. L., Russell, G. J., Gittleman, J. L. & Brooks, T. M. (1995) The future of biodiversity. *Science*, **269**, 347-350.

Polasky, S., Camm, J. D., Solow, A. R., Csuti, B., White, D., Ding, R. (2000) Choosing reserve networks with incomplete species information. *Biological Conservation*, **94**, 1-10.

Real, R., Barbosa, A. M. & Vargas, J. M. Obtaining environmental favourability functions from logistic regression. *Environmental and Ecological Statistics*, en prensa.

Reddy, S. & Dávalos, L. M. (2003) Geographical sampling bias and its implications for conservation priorities in Africa. *Journal of Biogeography*, **30**, 1719-1727.

Reineking, B. & Schröder, B. (2003) Computer-intensive methods in the analysis of species-habitat relationships. En *GfÖ Arbeitskreis Theorie in der Ökologie*, eds. H. Reuter, B. Breckling & A. Mittwollen, pp. 100-117. P. Lang Verlag Frankfurt/M.

Reutter, B. A., Helfer, V., Hirzel, A. H. & Vogel, P. (2003) Modelling habitat-suitability using museum collections: an example with three sympatric *Apodemus* species from the Alps. *Journal of Biogeography*, **30**, 581–590.

Richardson, B. J., Zabka, M., Gray, M. R. & Milledge, G. (2006) Distributional patterns of jumping spiders (Araneae: Salticidae) in Australia. *Journal of Biogeography*, **33**, 707-719.

Robertson, M. P., Villet, M. H. & Palmer, A. R. (2004) A fuzzy classification technique for predicting species´ distributions: applications using invasive alien plants and indigenous insects. *Diversity and Distributions*, **10**, 461-474.

Rojas, A. B., Cotilla, I., Real, R. & Palomo, L. J. (2001) Determinación de las áreas probables de distribución de los mamíferos terrestres en la provincia de Málaga. *Galemys*, **13**, 217-229.

Samways, M. J. (2005) *Insect diversity conservation*. Cambridge University Press, Cambridge.

Sánchez-Cordero, V., Cirelli, V., Munguía, M. & Sarkar, S. (2005) Place prioritization for biodiversity representation using species´ ecological niche modeling. *Biodiversity Informatics*, **2**, 11-23.

Scott, J. M., Heglund, P. J., Haufler, J. B., Morrison, M., Raphael, M. G., Wall, W. B. & Samson, F. (Eds.) (2002) *Predicting species occurrences: Issues of accuracy and scale.* Island Press, Covelo, California.

Segurado, P. & Araújo, M. B. (2004) An evaluation of methods for modelling species distributions. *Journal of Biogeography*, **31**, 1555-1568.

Seoane, J., Bustamante, J. & Díaz-Delgado, R. (2005) Effect of expert opinion on the predictive ability of environmental models of bird distribution. *Conservation Biology*, **19**, 512-522.

Seoane, J., Justribó, J. H., García, F., Retamar, J., Rabadán, C. & Atienza, J. C. (2006) Habitat-suitability modelling to assess the effects of land-use changes on Dupont´s lark *Chersophilus duponti*: A case study in the Layna Important Bird Area. *Biological Conservation*, **128**, 241-252.

Soberón, J. & Llorente, J. (1993) The use of species accumulation functions for the prediction of species richness. *Conservation Biology*, **7,** 480-488.

Soberón, J. & Peterson, A. T. (2005) Interpretation of models of fundamental ecological niches and species' distributional areas. *Biodiversity Informatics*, **2**, 1-10.

Steyerberg, E. W., Eijkemans, M. J. C. & Habbema, J. D. F. (1999) Stepwise selection in small data sets: a simulation study of bias in logistic regression analysis. *Journal of Clinical Epidemiology*, **52**, 935-942.

Stockwell, D. & Peters, D. (1999) The GARP modeling system: problems and solutions to automated spatial prediction. *International Journal of Geographic Information Science*, **13**, 143-158.

Stockwell, D. R. B. & Peterson, A. T. (2002) Effects of sample size on accuracy of species distribution models. *Ecological Modelling*, **148**, 1-13.

Stork, N. E. (1997) Measuring global biodiversity and its decline. En *Biodiversity II: Understanding and Protecting our Biological Resources*, eds. M. L. Reaka-Kudla, D. O. Wilson & E. O. Wilson, pp. 41-68. Joseph Henry Press, Washington D. C.

Termansen, M., McClean, C. J. & Preston, C. D. (2006) The use of genetic algorithms and Bayesian classification to model species distributions. *Ecological Modelling*, **192**, 410-424.

Thuiller, W., Araújo, M. B. & Lavorel, S. (2003) Generalized models vs. classification tree analysis: predicting spatial distributions of plant species at different scales. *Journal of Vegetation Science*, **14**, 669-680.

Thuiller, W., Lavorel, S., Araújo, M. B., Sykes, M. T. & Prentice, I. C. (2005) Climate change threats to plant diversity in Europe. *Proceedings of the National Academy of Sciences, USA*, **102**, 8245-8250.

Tyre, A. J., Tenhumberg, B., Field, S. A., Niejalke, D., Paris, K. & Possingham, H. P. (2003) Improving precision and reducing bias in biological surveys by estimating false negative error rates in presence-absence data. *Ecological Applications*, **13**, 1790–1801.

van Jaarsveld, A. S., Freitag, S., Chown, S. L., Muller, C., Koch, S., Hull, H., Bellamy, C., Krüger, M., Endrödy-Younga, S., Mansell, M. W. & Scholtz, C. H. (1998) Biodiversity assessment and conservation strategies. *Science*, **279**, 2106-2108.

Van Nes, E. H. & Scheffer, M. (2005) A strategy to improve the contribution of complex simulation models to ecological theory. *Ecological Modelling*, **185**, 153-164.

Vaughan, I. P. & Ormerod, S. J. (2003) Improving the quality of distribution models for conservation by addressing shortcomings in the field collection of training data. *Conservation Biology*, **17**, 1601-1611.

Wiens, J. J. & Graham, C. H. (2005) Niche conservatism: integrating evolution, ecology, and conservation biology. *Annual Review of Ecology, Evolution and Systematics*, **36**, 519-539.

Wilson, E. O. (1992) *La diversidad de la vida.* Crítica, Barcelona.

Yoccoz, N. G., Nichols, J. D. & Boulinier, T. (2001) Monitoring of biological diversity in space and time. *Trends in Ecology and Evolution*, **16**, 446-453.

Zaniewski, A. E., Lehmann, A. & Overton, J. McC. (2002) Predicting species spatial distributions using presence-only data: a case study of native New Zealand ferns. *Ecological Modelling*, **157**, 261-280.

Zimmermann, N. E. & Kienast, F. (1999) Predictive mapping of alpine grasslands in Switzerland: species versus community approach. *Journal of Vegetation Science*, **10**, 469-482.

# Primera Parte: Riqueza de las familias Araneidae y Thomisidae en la Comunidad de Madrid

# UN PROTOCOLO DE MUESTREO COMBINADO PARA LA ESTIMACIÓN DE LOS ENSAMBLAJES DE ARANEIDAE Y THOMISIDAE (ARACHNIDA, ARANEAE)

**RESUMEN.** A medida que se acelera la desaparición de las especies se hace más urgente el desarrollo de protocolos de muestreo basados en métodos eficientes. El conocimiento de la aracnofauna ibérica es bastante escaso, por lo que es necesario llevar a cabo inventarios fiables y tan completos como sea posible, de una manera rápida y sencilla. En el presente trabajo se comparan seis métodos diferentes de muestreo (mangueo, batido, trampas de interceptación, captura directa a dos alturas distintas y análisis de hojarasca) para el inventariado de las familias Araneidae y Thomisidae en parcelas de 1 $km^2$, estudiando su comportamiento en tres hábitats con diferente complejidad estructural de la vegetación. Los resultados muestran que, para conseguir inventarios fiables de estas dos familias, es necesaria la combinación de mangueo, batido y de las trampas de caída. En los hábitats en los que la localización de los araneidos es sencilla debido a que se concentran en parches de vegetación concretos, la captura directa a una altura por encima de las rodillas contribuye a mejorar el protocolo.

**Palabras clave:** inventarios de riqueza específica, métodos de muestreo, eficiencia, complementariedad

**Este capítulo ha sido publicado en:**

JIMÉNEZ-VALVERDE, A. & LOBO, J. M. (2005). Determining a combined sampling procedure for a reliable estimation of Araneidae and Thomisidae assemblages (Arachnida: Araneae). *Journal of Arachnology*, **33**, 33-42.

# DETERMINING A COMBINED SAMPLING PROCEDURE FOR A RELIABLE ESTIMATION OF ARANEIDAE AND THOMISIDAE ASSEMBLAGES (ARACHNIDA, ARANEAE)

**ABSTRACT.** As the disappearance of species accelerates, it becomes extremely urgent to develop sampling protocols based on efficient sampling methods. As knowledge of the Iberian spider fauna is extremely incomplete, it is becoming necessary to facilitates reliable and complete species richness inventory collection. In this work the results from six sampling methods (sweeping, beating, pitfall traps, hand collecting at two different heights and leaf litter analysis) in three habitats with different vegetation structure are compared for the inventory of Araneidae and Thomisidae in 1 $km^2$ sampling plots. A combination of sweeping, beating and pitfall trapping prove to be necessary to achieve a reliable inventory of these two spider families. Hand collecting above knee level contributes to the improvement of the protocol in certain habitats where araneids, concentrated in patches of suitable vegetation, are easy to find.

**Keywords:** species richness inventory, sampling methods, efficiency, complementarity

## INTRODUCTION

Loss of biodiversity, one of the greatest environmental problems (Wilson, 1988; May *et al.*, 1995), the outcome of the accelerating destruction of ecosystems, means that many species will be eradicated while still undiscovered or unstudied. Protecting biodiversity implies protecting terrestrial arthropods, a group poorly known but

comprising around 80% of the Earth's species and including those denominated as hiperdiverse (Hammond, 1992). Those groups are the least understood, yet contribut most to the planet's biotic diversity. Conservation of biological diversity requires detailed information on the geographic distribution of organisms. In the case of arthropods, as this information is almost impossible to acquire in the medium-term by means of field sampling (Ehrlich & Wilson, 1991; Williams & Gaston, 1994), the utilization of predictive model techniques may be the only possible way to estimate the distribution of biodiversity attributes (Margules *et al.*, 1987; Iverson & Prasad, 1998; Guisan & Zimmermann, 2000; Lobo & Martín-Piera, 2002; etc). However, application of these predictive methods requires reliable biological information; when this is lacking, the design of specific sampling protocols for each taxonomic group that gather the maximum information, most cost-effectively, becomes essential.

About 36000 species of the order Araneae have been described, while the total number is estimated at between 60000 and 170000 (Coddington & Levi, 1991; Platnick, 1999). This is one of the most diversified orders (Coddington & Levi, 1991) and offers the greatest potential to help regulate terrestrial arthropod populations (Marc *et al.* 1999). Araneids, one of the most successful spider families (approximately 2600 species; Foelix, 1996), are relatively easy to detect due to their size, coloration, and their orb webs. Vegetation structure seems to be the most important parameter in determining their presence (Wise, 1993). Unlike the araneids, thomisids (crab spiders) do not use webs to capture prey; instead they ambush prey from flowers or leaves (Wise, 1993), where their cryptic coloration allows them to go unnoticed. Some genera, like *Xysticus* and *Ozyptila*, are eminently edaphic, capturing prey among leaf litter and herbaceous vegetation.

Arachnological tradition is sorely lacking in the Iberian Peninsula, and spider distribution is extremely poorly understood (1180 recorded species; Morano, 2002). Only in the province of Aragón is there a recent catalogue of arachnological fauna (Melic, 2000); the rest of the Iberian catalogues include out-dated records, most of doubtful quality and erroneous (Melic, 2001). So, it is necessary to augment taxonomic and distributional data on Iberian spiders by using effective and standardized sampling protocols, the design of which involves overcoming some difficulties. As spiders' life history, behavior and morphologic, physiological and ecological adaptation vary widely (Turnbull, 1973), sampling method effectiveness depends on the nature of the taxonomic group (Canard, 1981; Churchill, 1993; Coddington *et al.*, 1996; Costello & Daane, 1997; Churchill & Arthur, 1999). Furthermore, it must be kept in mind that the effectiveness of the method also depends on the environment (Canard, 1981). Thus, in order to inventory reliably and completely, the design of the sampling protocol should combine various sampling methods, selecting the methods promising maximum information and complementarity for each environment and taxonomic group (Coddington *et al*., 1996; Green, 1999; Sørensen *et al*., 2002). In this work, several sampling methods for Araneidae and Thomisidae species are compared, in habitats with distinct vegetation complexity, in order to determine which combination captures the maximum number of species with the minimum number of sampling techniques.

## METHODS

**Study site. —** The study was carried out from 2 May - 14 June 2002 in three localities in the Comunidad de Madrid (central Spain), with vegetation differing in structural complexity as follows: 1) A grassland zone subjected to intense pasturing

pressure, with small shrub patches, at 980 m elevation in the municipality of Colmenar Viejo (latitude 40.69, longitude -3.77). Its potential vegetation is the holm-oak forest (supra-mesomediterranean-siliceous series of *Quercus ilex rotundifolia*; Rivas-Martinez, 1987). 2) An extensive and dense zone of shrub located in El Berrueco (latitude 39.97, longitude -3.53), at 940 m elevation. The area belongs to the same vegetation series as the former (Rivas-Martinez, 1987); nevertheless, human activity has caused the original vegetation to be replaced by the *Cistus ladanifer* series, with patches of *Lavandula pedunculata* and *Thymus* spp. 3) A Holm-oak forest zone in Cantoblanco (latitude 40.51, longitude -3.65) at an elevation of 700 m, composed of some tall (6-8 m) specimens of *Quercus ilex rotundifolia*, though the majority of the trees are between 3-4 m tall. An old plantation of *Pinus pinea*, which dates from the 1930s, occupies one part of the forest.

**Sampling methods. —-** In each habitat a 1 $km^2$ sampling plot divided into 2500 subplots of 400 $m^2$ was established; 20 of these subplots were chosen at random, and a sampling effort unit carried out in each. For the capture of species in these two families, six cheap, easy and widely used sampling methods were employed: sweeping, beating, pitfall traps, above-knee-level visual search, below-knee-level visual search, and leaf litter analysis. A sampling effort unit was defined as one of the following: 1) A one-person sweep of the herbaceous vegetation and shrub during 15 minutes. The opening of the sweep net was 37 cm in diameter, and it was emptied at regular intervals to avoid loss and destruction of the specimens. 2) A one-person beating of bushes and small trees and branches during 15 minutes with a heavy stick; the specimens fell on a 1.25 × 1.25 m white sheet. In cases where the structure of the vegetation made the use of the sheet difficult a 41 × 29 cm plastic pail was employed. 3) A one-person visual search from knee level to as high as one can reach (above visual search, AVS) during 15 minutes. 4)

A one-person visual search from ground to knee level (below visual search, BVS) during 15 min. Stones were lifted up because thomisids, especially females after laying eggs (Levy, 1975; Hidalgo, 1986), from the genera *Xysticus* and *Ozyptila* usually dwell under them. 5) Analysis during 15 min. of leaf litter poured in a white pail, justifiable because this is the habitat of the generus *Ozyptila* (Thomisidae) (Urones, 1998). 6) The running of 4 open pitfall traps during 48 hours. These traps were 11.5 cm wide and 1 liter in volume, each 10 m apart from the others in order to avoid interference effects and to maximize the efficacy of each trap (Samu & Lövei, 1995). Traps were filled with water, and a few drops of detergent added to break the surface tension so as to prevent the spiders from escaping.

Spiders were sucked up with a pooter to reduce damage and were transferred to 70% alcohol. Sampling was always done by the same person in order to avoid possible differences due to the effect of the collector (Norris, 1999); rainy and windy days were avoided in order to prevent a reduction in the efficiency of the sampling methods (see Gyenge *et al*., 1997 and Churchill & Arthur, 1999). All specimens are deposited in the Museo Nacional de Ciencias Naturales collection (Madrid, Spain). All together, sampling involved running 240 pitfall traps (3 sampling plots × 20 subplots × 4 pitfall traps) and one-person fieldwork during 75 hours (0.25 hours × 5 methods × 3 sampling plots × 20 subplots).

**Data analysis. —** The cumulative number of species found by different sampling efforts (species accumulation curves) was studied to evaluate the accuracy of the species inventories obtained in each of the three sampling plots (see Gotelli & Colwell, 2001). The number of sampling effort units (i.e. the number of subplots) was used as the measure of sampling effort, and the order in which sampling unit inventories were added was randomized 500 times to build smoothed curves using the EstimateS

5.0.1 software (Colwell, 1997). The asymptotic value of the accumulation curves obtained was estimated using the Clench equation (Soberon & Llorente, 1993; Colwell & Coddington, 1994). This score, together with the species richness estimations produced by three nonparametric methods, was used to test if the total number of species caught in each sampling plot underestimated the true species richness. The nonparametric species richness estimators used are the first-order jackknife, the abundance-based coverage (ACE), and the incidence-based coverage estimator (ICE). Detailed descriptions of the estimators can be found in Colwell (1997) and Colwell & Coddington (1994).

In order to study the effects of sampling method and the interaction of method and habitat on the number of species and individuals collected per sampling effort unit, a factorial ANOVA was performed. As data were not normally distributed, they were transformed by $log(n+1)$, and a Tukey test (HDS) was used to determine pairwise significant differences ($p < 0.05$). STATISTICA package (StatSoft, 2001) was used for all statistical computations.

**Other methodological considerations. —** As Norris (1999) pointed out, the inclusion of immature specimens is the factor which has the most significant effect on community trends. It cannot be assumed that the abundance distribution of juveniles is the same as that for adults, and the relative abundance of species in a community can be highly altered if juveniles are considered. However, since our objective was to find all the species inhabiting the sampling plots, juveniles that could be identified to the species level were included in the analysis. Sometimes genera represented only by immature states did appear, in which case, they were also included. Rejecting juveniles would have involved rejecting valuable information, and as they increased sample sizes significantly, their inclusion allowed statistical analysis. In araneids and thomisids,

unlike in most other spider families, color and morphology facilitate the identification of some juveniles. All together, 942 individuals were captured, 56% of them juveniles; almost half (247 individuals) have been used in the analysis.

## RESULTS

In 80 sampling effort units, a total of 661 individuals were captured, representing 26 species, 11 araneids and 15 tomisids.

**Completeness of the inventories. —** The Clench model function fits the accumulation curves well in each of the three sampling plots, with percentages of explained variation higher than 99% (Table 1 & Fig. 1). The predicted asymptote score does not differ too much from the observed species richness, the percentages of collected species oscillating around 80%. The nonparametric estimators used indicate that the collected species richness varies from 86% - 95% for the forest plot, 74% - 80% for the shrub plot, and 84% - 90% for the grassland plot. These results suggest that the exhaustiveness of the sampling in each of the three habitats is similar, so sampling plot composition and richness figures are comparable. However, still more intensive sampling should be necessary to obtain an accurate inventory in each habitat.

Table 1.- Observed species richness ($S_{obs}$) and results of four species richness estimators for each habitat. The relationship between the number of sampling effort units and the number of species was fitted to the asymptotic Clench equation (Colwell & Coddington, 1994) where $a/b$ is the asymptote and $R^2$ the percentage of explained variance. Jackknife 1 (first-order jackknife), ACE (abundance base coverage) and ICE (incidence-based coverage) are nonparametric estimators of species richness (Colwell, 1997).

| | Forest | Shrub | Grassland |
|---|---|---|---|
| $S_{obs}$ | 17 | 20 | 15 |
| Clench | $a/b$=21.6; $R^2$ = 99.9 | $a/b$=25.0; $R^2$ = 99.4 | $a/b$=18.8; $R^2$ = 99.9 |
| Jackknife 1 | 19.85 | 26.65 | 17.85 |
| ICE | 18.49 | 27.18 | 16.73 |
| ACE | 17.85 | 25.07 | 17.26 |

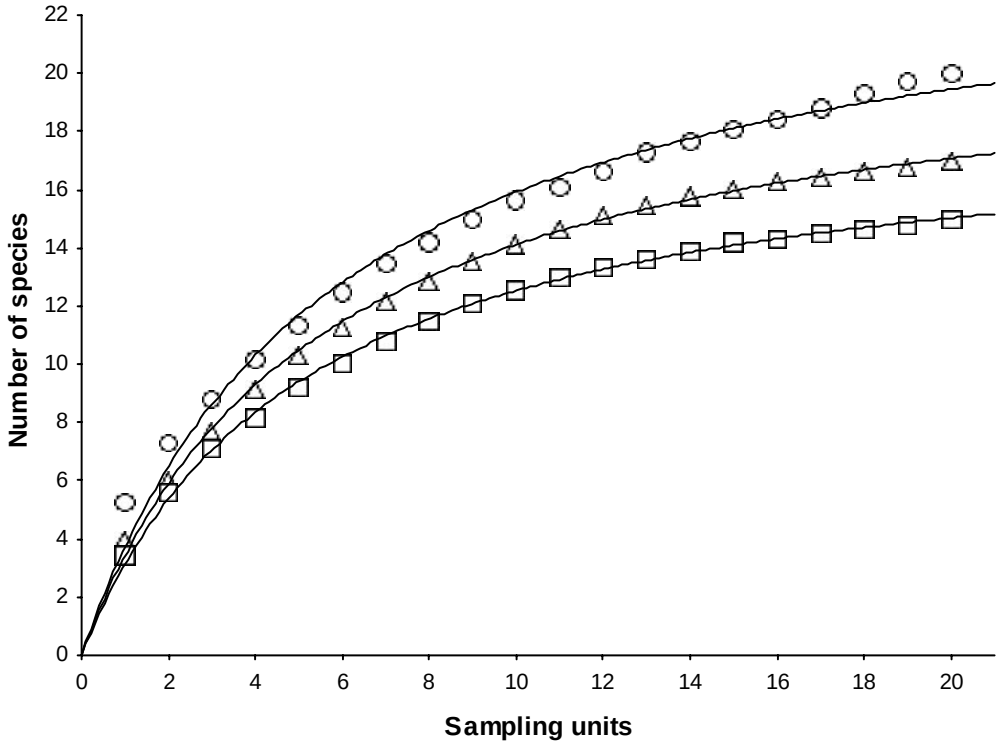



Figure 1.- Species accumulation curves for the three sampling plots with Clench function fitted: □ grassland; ○ shrub; Δ forest. The cumulative number of species found at different numbers of sampling effort units was randomized 500 times using the EstimateS 5.0.1 software (Colwell, 1997).

**Sampling method performance. —** From the three sampling plots, only one individual of *M. acalypha* (Walckenaer 1802) was captured by leaf litter analysis method (in the shrub plot). As this species was collected plentifully with the other sampling methods, the results of this technique are not considered. Both the mean number of collected species ($F_{(4,285)} = 58.5$; $p < 0.0001$) and the mean number of individuals ($F_{(4,285)} = 79.9$; $p < 0.0001$) differ statistically from one sampling method to another. Both in the case of species richness and for the number of individuals, all pairwise comparisons between sampling methods are significant by a posteriori Tukey HSD test, except in the case of pitfall traps and BVS, and beating and AVS (see Table 2). Sweeping, the technique which captured the greatest number of species and

individuals, with araneids making up 47 % of the species and 68% of the individuals collected, is also the method that captured more species not captured by any other sampling method (unique species, two araneids and three thomisids). Pitfall traps and BVS are the methods that captured the smallest number of species and individuals, but while BVS did not yield unique species, pitfall traps did capture two unique species. With pitfall traps, only thomisids of the genera *Xysticus* and *Ozyptila* were captured. In the case of the BVS, araneids make up 57% of the species and 62% of the individuals. With regard to the other sampling methods, beating and AVS yield the same number of individuals, though the total number of species is larger for the former. In beating, araneids make up 47% of the species and 43% of the individuals; using AVS araneid, captures were more frequent, accounting for 78% of species and 89% of individuals. AVS did not yield any unique species, while beating produced three unique thomisids.

Table 2.- Total number of individuals (N), mean number of individuals (±SE) per sampling unit ($N_{MEAN}$ ), total number of species (S), mean number of species(±SE) per sampling unit ($S_{MEAN}$ ), and number of unique species ($S_{UNI}$) for each sampling plot and each sampling method.

| | **Sampling Plot** | | |
|---|---|---|---|
| | **Forest** | **Shrub** | **Grassland** |
| N | 205 | 348 | 108 |
| $N_{MEAN}$ | 2.07±0.36 | 3.48±0.56 | 1.57±0.5 |
| S | 17 | 20 | 15 |
| $S_{MEAN}$ | 0.92±0.14 | 1.5±0.17 | 0.72±0.1 |

| | **Sampling Method** | | | | |
|---|---|---|---|---|---|
| | **Pitfall** | **Sweeping** | **Beating** | **BVS** | **AVS** |
| N | 25 | 442 | 90 | 13 | 91 |
| $N_{MEAN}$ | 0.41±0.14 | 8.08±1.06 | 1.6±0.23 | 0.22±0.09 | 1.55±0.28 |
| S | 5 | 17 | 15 | 7 | 9 |
| $S_{MEAN}$ | 0.3±0.08 | 2.7±0.24 | 1.08±0.14 | 0.18±0.07 | 0.98±0.14 |
| $S_{UNI}$ | 2 | 5 | 3 | 0 | 0 |

By an iterative procedure the sampling methods were ranked sequentially, for each habitat, according to contribution to total species richness in this habitat. Both in the forest and shrub, sweeping is the method that yielded more species, followed by beating and pitfall traps. Together, these three methods captured all the observed species in these habitats. In grassland, where a broader combination of methods is necessary to obtain a reliable inventory (Table 3), beating captured more species, while sweeping, AVS and pitfall traps or BVS seem to be indispensable.

Table 3.- Results of a complementarity procedure in which the inventories of each sampling method were sequentially selected for each habitat according to its contribution to the species richness.

| Habitat | Iteration | Sampling method | Number of species | Accumulated species |
|---|---|---|---|---|
| Forest | 1 | Sweeping | 12 | 12 |
| | 2 | Beating | 4 | 16 |
| | 3 | Pitfall | 1 | 17 |
| Shrub | 1 | Sweeping | 13 | 13 |
| | 2 | Beating | 4 | 17 |
| | 3 | Pitfall | 3 | 20 |
| Grassland | 1 | Beating | 8 | 8 |
| | 2 | Sweeping | 4 | 12 |
| | 3 | AVS | 2 | 14 |
| | 4 | Pitfall or BVS | 1 | 15 |

**Sampling method-habitat interaction. —** The mean number of species per sampling unit ($F_{(2, 285)}$=15.14; $p < 0.001$), as well as the mean number of individuals ($F_{(2, 285)}$=15.73; $p < 0.001$), differs significantly between sampling plots. According to a posteriori Tukey HDS test, only in the shrub sampling plot is the species richness and number of individuals significantly greater than in the other two sampling plots (Table 2). However, sampling method and habitat interaction significantly affect both the mean number of species ($F_{(8, 285)} = 6.6$; $p < 0.0001$) and the mean number of individuals

per sampling unit ($F_{(8,\ 285)} = 9.6$; $p < 0.0001$), indicating that the performance of the various sampling methods depends on the habitat.

The results of a posteriori Tukey HSD test highlight the significantly different interaction terms. The scheme generated for the mean number of species and individuals is quite similar (Fig. 2). There is not a significant between-habitat variation in the number of individuals or species collected by pitfall-traps, BVS or beating. The AVS method collected a significantly greater number of species and individuals in shrub and grassland than in forest (Fig. 2), only in the grasslands did it capture more species and individuals than BVS and pitfall traps; its captures equalled those of beating in the three habitats. Likewise, sweeping method captures also varied with habitat; the mean number of species and individuals captured in grasslands was significantly smaller than in the other two habitats (Fig. 2). Indeed, the sweeping method captured more species and individuals in forest and shrub, while in grassland its performance was similar to that of beating or AVS.

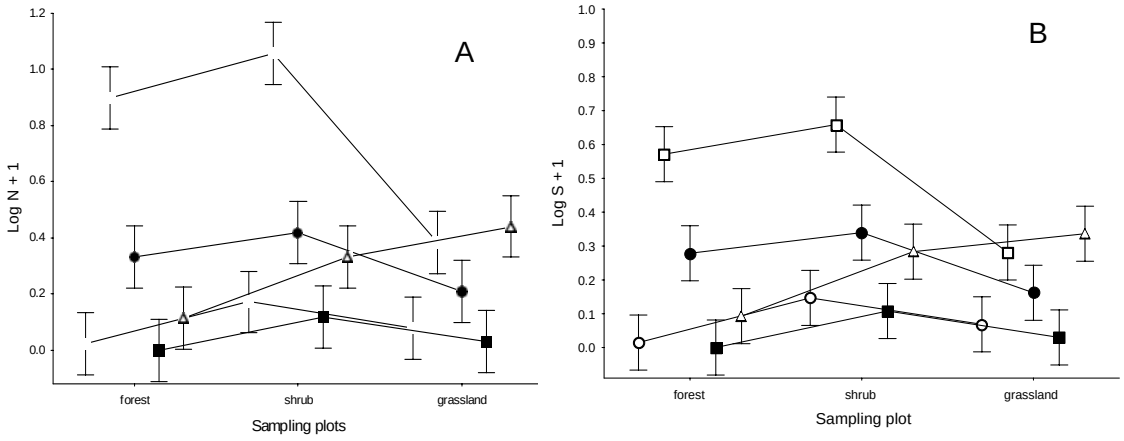



Figure 2.- Variation in the mean number of individuals (log of N + 1; ±95% confidence interval) per sample (A) and mean number of species (log of S + 1; ±95% confidence interval) per sample (B) between the three studied habitats or sampling plots.□, sweeping;●, beating; Δ, AVS; ■, BVS; ○, pitfall trapping.

## DISCUSSION

Methods differ greatly in the number of species and individuals caught, and its performance depends on vegetation structure. Sweeping is a standard item in arachnologists´ fieldwork due to its ease of use and effectiveness (Buffington & Redak, 1998). It was the most efficient sampling method in forest and shrub sampling plots, as it was the one which yielded more species and individuals. However, in the grassland sampling plot, the extreme shortness of the grass and the presence of thorny shrub patches limited its use; AVS and beating there prodeced equal value of mean individuals and species richness. While other authors have also noticed the reduced usefulness of sweeping in certain habitats (Churchill & Arthur, 1999), as sweeping was found here to yield unique species in the three sampling plots, it must continue to be fundamental to sampling protocol.

Because beating and AVS work on similar vegetation habitats, they sample the same part of the spider community. However, while beating yielded unique species in the three habitats, AVS only did so in the grassland sampling plot, where araneids were concentrated in shrub patches and therefore easily spotted. Furthermore, AVS, a sampling method biased towards big and flashy spiders, yielded a greater proportion of araneids. It can be noticed that where vegetation structure makes visual search difficult, i.e. in the forest sampling plot, AVS is less efficient and beating yielded more (although not statiscally significant) species and individuals. Beating must be added to the sampling protocol, along with AVS in habitats with such a vegetation structure that the visual detection of individuals is easy.

Although its efficiency was quite low in our study, pitfall trapping, one of the most frequently used methods to sample surface-active terrestrial arthropod communities, is essential for sampling that part of the assemblage (i.e., genera *Xysticus*

and *Ozyptila*, which comprise more than the 70% of the Iberian thomisid fauna). Indeed, all the pitfall captures in the three sampling plots belong to these two genera. As already noted by other authors (Churchill, 1993; Standen, 2000), the captures of this sampling method were biased in favor of adult individuals, facilitating the identification of the specimens and helping in the inventory work. As BVS samples the same part of the community as pitfall traps do and does not contribute unique species, it can be done without. Thus, only pitfall trapping must be included in the sampling protocol.

Because the aim of this sampling protocol is the estimation of species richness, visual search could be more efficient if centered on new species, ignoring the common ones (Dobyns, 1997; Churchill & Arthur, 1999). The paucity of species and individuals captured by pitfall trapping suggests that the inventory would have been more effective if greater sampling effort were allocated to this method. Brennan *et al*. (1999) found that the larger the pitfall trap diameter, the greater the number of species captured. Work *et al*. (2002) pointed out that larger traps were more effective in the characterization of rare elements of an epigeal fauna. They also recommended combining large traps with smaller ones in order to sample a greater range of microhabitats. However, it is difficult to judge if the protocol would have been improved by changing the pitfall trap design or by trying another method that samples this epigeal fauna more accurately.

For none of the three sampling sites does the observed species accumulation curve reach an asymptote, although it seems that the simpler the vegetation structure, the smaller the curve-asymptote separation, and the smaller the difference between $S_{obs}$ and the Clench model estimation from the nonparametric estimator values. Tight clustering of these three nonparametric estimators was also found by Toti *et al*. (2000), suggesting that they either estimate the same real value or are biased similarly. Other researchers working with the entire spider fauna (Coddington *et al*., 1996; Dobyns,

1997; Toti *et al*., 2000; Sørensen *et al*., 2002) have also failed to produce asymptotic species accumulation curves. However, according to the estimations obtained, the three inventories sampled around 80% of spider fauna, indicating that it is possible to estimate the probable number of species in a 1 $km^2$ plot. The percentages of completeness are quite similar to those found by other authors in temperate forests (Dobyns, 1997, 89%; Sørensen *et al*., 2002, 86-89%).

Our study is just a spring "snapshot" of the entire annual spider species richness of three sampling plots in different habitats. Spider assemblages, dynamic during the season, change in species composition. Thus, results depend on the time of sampling (Churchill & Arthur, 1999; Riecken, 1999). Nevertheless, estimating species richness accurately at a given time carries weight because sampling designs for annual studies depend on it (Coddington *et al*., 1996; Sørensen *et al*., 2002). Determining the proportion of the entire annual spider fauna that is represented in the spring sample is an objective of work currently being carried out.

Spider life history and behavioral diversity pose a challenge to the development of a precise and cost-effective sampling program (Costello & Daane, 1997). Studies that have tried to take in the entire range of spider fauna have found that even intensive sampling does not reflect the whole of species richness (Coddington *et al*., 1996; Toti *et al*., 2000; Sørensen *et al*., 2002). So, Sørensen *et al*. (2002) suggest that long-term monitoring programs should focus on single, or few, families, or a single feeding guild, and use a few standardized and practical sampling methods. Our study has focused on two abundant spider families, Araneidae and Thomisidae, and has shown that a particular combination of sampling methods in each habitat is required to optimize efficacy and minimize effort. Sweeping, beating, pitfall traps and AVS in specific locations yield a reliable inventory of these two spider taxa in a 1 $km^2$ plot. Given how

imperative a more detailed knowledge of Iberian spiders is, additional studies should be carried out in order to develop standardized sampling protocols for other spider families and/or guilds.

## ACKNOWLEDGMENTS

This paper has been supported by the project "Faunística Predictiva: Análisis comparado de la efectividad de distintas metodologías y su aplicación para la selección de reservas naturales" (grant: REN 2001-1136/GLO), and also by a PhD Museo Nacional de Ciencias Naturales/C.S.I.C./Comunidad de Madrid grant.

## LITERATURE CITED

Brennan, K. E. C., Majer, J. D. & Reygaert, N. (1999) Determination of an optimal pitfall trap size for sampling spiders in Western Australian Jarrah forest. *Journal of Insect Conservation*, **3**, 297-307.

Buffington, M. L. & Redak, R. A. (1998) A comparison of vacuum sampling versus sweep-netting for arthropod biodiversity measurements in California coastal sage scrub. *Journal of Insect Conservation*, **2**, 99-106.

Canard, A. (1981) Utilisation comparée de quelques méthodes d´échantillonnage pour l´étude de la distribution des araignées en landes. *Atti della Società Toscana di Scienze Naturali. Serie B.*, **Memorie 88, suppl.**, 84-94.

Churchill, T. B. (1993) Effects of sampling method on composition of a Tasmanian coastal heathland spider assemblage. *Memoirs of the Queensland Museum*, **33(2)**, 475-481.

Churchill, T. B. & Arthur, J. M. (1999) Measuring spider richness: effects of different sampling methods and spatial and temporal scales. *Journal of Insect Conservation*, **3**, 287-295.

Coddington, J. A. & Levi, H. W. (1991) Systematics and evolution of spiders. *Annual Review of Ecology and Systematics*, **22**, 565-592.

Coddington, J. A., Young, L. H. & Coyle, F. A. (1996) Estimating spider species richness in a southern Appalachian cove hardwood forest. *Journal of Arachnology*, **24,** 111-128.

Colwell, R. K. (1997) *EstimateS: Statistical Estimation of Species Richness and Shared Species from Samples (Software and User´s Guide), Version 5.0.1*, available in http://viceroy.eeb.uconn.edu/estimates

Colwell, R. K. & Coddington, J. A. (1994) Estimating terrestrial biodiversity through extrapolation. *Philosophical Transactions of the Royal Society (series B)*, **345**, 101-118.

Costello, M. J. & Daane, K. M. (1997) Comparison of sampling methods used to estimate spider (Araneae) species abundance and composition in grape vineyards. *Environmental Entomology*, **26(2)**, 142-149.

Dobyns, J. R. (1997) Effects of sampling intensity on the collection of spider (Araneae) species and the estimation of species richness. *Environmental Entomology*, **26(2)**, 150-162.

Ehrlich, P. R. & Wilson, E. O. (1991) Biodiversity studies: science and policy. *Science*, **253**, 750-752.

Foelix, R. F. (1996) *Biology of spiders*. Oxford University Press, New York.

Gotelli, N. J. & Collwell, R. K. (2001) Quantifying biodiversity: procedures and pitfalls in the measurement and comparison of species richness. *Ecology Letters,* **4**, 379-391.

Green, J. (1999) Sampling method and time determines composition of spider collections. *Journal of Arachnology*, **27**, 176-182.

Guisan, A. & Zimmermann, N. E. (2000) Predictive habitat distribution models in ecology. *Ecological Modelling,* **135**, 147-186.

Gyenge, J. E., Edelstein, J. D. & Trumper, E. V. (1997) Comparación de técnicas de muestreo de artrópodos depredadores en alfalfa y efecto de factores ambientales sobre sus estimaciones de abundancia. *Ceiba*, **38(1)**, 13-18.

Hammond, P. M. (1992) Species inventory. In *Global biodiversity, status of the earth´s living resources*, ed. B. Groombridge, pp. 17-39. Chapman & Hall, London.

Hidalgo, I. L. (1986) *Estudio de los tomísidos de la provincia de León (Araneae: Thomisidae & Philodromidae)*. Excma. Diputación Provincial de León. Inst. "Fray Bernardino de Sahagún".

Iverson, L. R. & Prasad, A. M. (1998) Predicting abundance of 80 tree species following climate change in the eastern United States. *Ecological Monographs*, **68**, 465-485.

Levy, G. (1975) The spider genera *Synaema* and *Ozyptila* in Israel (Araneae: Thomisidae). *Israel Journal of Zoology*, **24**, 155-175.

Lobo, J. M. & Martín-Piera, F. (2002) Searching for a predictive model for Iberian dung beetle species richness (Col., Scarabaeinae) using spatial and environmental variables. *Conservation Biology*, **16**, 158-173.

Marc, P., Canard, A. & Ysnel, F. (1999) Spiders (Araneae) useful for pest limitation and bioindication. *Agriculture, Ecosystems and Environment*, **74**, 229-273.

Margules, C. R., Nicholls, A. O. & Austin, M. P. (1987) Diversity of *Eucalyptus* species predicted by a multi-variable environment gradient. *Oecologia*, **71**, 229-232.

May, R. M., Lawton, J. H. & Stork, N. E. (1995) Assessing extinction rates. In *Extinction Rates*, eds. J. H. Lawton & R. M. May, pp. 1-24. Oxford University Press, Oxford.

Melic, A. (2000) Arañas de Aragón (Arachnida: Araneae). *Catalogus de la Entomofauna Aragonesa*, **22**, 3-40.

Melic, A. (2001) Arañas endémicas de la Península Ibérica e Islas Baleares (Arachnida: Araneae). *Revista Ibérica de Aracnología*, **4**, 35-92.

Morano, E. (2002) *Catálogo Ibérico de Arañas,* in http://entomologia.rediris.es/gia/catalogo/index.htm

Norris, K. C. (1999) Quantifying change through time in spider assemblages: sampling methods, indices and sources of error. *Journal of Insect Conservation*, **3**, 309-325.

Platnick, N. I. (1999) Dimensions of biodiversity: Targeting megadiverse groups. In *The living Planet in Crisis: Biodiversity Science and Policy*, eds. J. Cracraft & F. T. Grifo, pp. 33-52. Columbia Univ. Press, New York.

Riecken, U. (1999) Effects of short-term sampling on ecological characterization and evaluation of epigeic spider communities and their habitats for site assessment studies. *Journal of Arachnology*, **27**, 189-195.

Rivas-Martínez, S. (1987) *Memoria del mapa de series de vegetación de España 1:400.000*. ICONA, Madrid.

Samu, F. & Lövei, G. L. (1995) Species richness of a spider community (Araneae): extrapolation from simulated increasing sampling effort. *European Journal of Entomology*, **92**, 633-638.

Soberón, J. & Llorente, B. J. (1993) The use of species accumulation functions for the prediction of species richness. *Conservation Biology*, **7**, 480-488.

Sørensen, L. L., Coddington, J. A. & Scharff, N. (2002) Inventorying and estimating subcanopy spider diversity using semiquantitative sampling methods in an afromontane forest. *Environmental Entomology*, **31**, 319-330.

Standen, V. (2000) The adequacy of collecting techniques for estimating species richness of grassland invertebrates. *Journal of Applied Ecology*, **37**, 884-893.

StatSoft (2001) *STATISTICA (data analysis software system and user´s manual). Version 6.* StatSoft, Inc., Tulsa, OK.

Toti, D. S., Coyle, F. A & Miller, J. A. (2000) A structured inventory of Appalachian grass bald and heath bald spider assemblages and a test of species richness estimator performance. *Journal of Arachnology*, **28**, 329-345.

Turnbull, A. L. (1973) Ecology of true spiders (Araneomorphae). *Annual Review of Entomology,* **18**, 305-348.

Urones, C. (1998) Descripción de *Oxyptila bejarana* n. sp. de la Sierra de Béjar (Salamanca, España) (Araneae, Thomisidae). *Revue Arachnologique*, **12(8)**, 79-88.

Williams, P. H. & Gaston, K. J. (1994) Measuring more of biodiversity: can higher-taxon richness predict wholesale species richness? *Biological Conservation*, **67**, 211-217.

Wilson, E. O. (1988) The current state of biological diversity. In *Biodiversity,* ed. E. O. Wilson, pp. 3-17. National Academic Press, Washington, D. C.

Wise, D. H. (1993) *Spiders in ecological webs*. Cambridge University Press, New York.

Work, T. T., Buddle, C. M., Korinus, L. M. & Spence, J. R. (2002) Pitfall trap size and capture of three taxa of litter-dwelling arthropods: implications for biodiversity studies. *Environmental Entomology*, **31(3)**, 438-448.

# DEFINIENDO PROTOCOLOS DE MUESTREO ÓPTIMOS DE ARAÑAS (ARANEAE, ARANEIDAE Y THOMISIDAE): ESTIMACIÓN DE LA RIQUEZA ESPECÍFICA, COBERTURA ESTACIONAL Y CONTRIBUCIÓN DE LOS INDIVIDUOS JUVENILES A LA RIQUEZA Y COMPOSICIÓN DE ESPECIES

**RESUMEN.** Se estudia la capacidad de protocolos reducidos en el tiempo para muestrear de manera fiable los ensamblajes mediterráneos anuales de arañas (Araneidae y Thomisidae), así como la contribución de los individuos juveniles en la estimación de la riqueza específica. Un muestreo anual estandarizado en una cuadrícula de 1 $km^2$ en el centro de España proporcionó inventarios fiables de Araneidae y Thomisidae. Para comparar la eficiencia de diversos diseños de muestreo, se estimó el grado de representatividad de un "muestreo óptimo", aquel efectuado en el número mínimo de meses para colectar el número total de especies. Se estimó también el grado de representatividad de un muestreo primaveral, así como el de los muestreos mensuales. Los cálculos se efectuaron incluyendo y excluyendo a los individuos juveniles. Cuando se deben muestrear múltiples localidades y, a la vez, hay que minimizar el esfuerzo empleado en el trabajo de campo, un muestreo efectuado durante un mes de la época primaveral aporta buenas estimas de la fauna de primavera, permitiendo la comparación entre localidades durante esta época de mayor riqueza. Nuestros resultados indican que los individuos juveniles deben ser tenidos en cuenta con el fin de obtener estimas fiables de la riqueza de especies, y deben ser almacenados a parte con el fin de analizarlos a medida que se avance en su identificación.

**Palabras clave:** inventarios de especies, Araneae, Araneidae, Thomisidae, muestreos reducidos en el tiempo, juveniles, estimación de la riqueza específica

**Este capítulo ha sido publicado en:**

JIMÉNEZ-VALVERDE, A. & LOBO, J. M. (2006). Establishing reliable sampling protocols for spider (Araneae, Araneidae & Thomisidae) assemblages: estimation of species richness, seasonal coverage and effect of juvelines on species richness and composition. *Acta Oecologica*, en prensa.

# ESTABLISHING RELIABLE SPIDER (ARANEAE, ARANEIDAE & THOMISIDAE) ASSEMBLAGE SAMPLING PROTOCOLS: ESTIMATION OF SPECIES RICHNESS, SEASONAL COVERAGE AND CONTRIBUTION OF JUVENILE DATA TO SPECIES RICHNESS AND COMPOSITION

**ABSTRACT.** The capacity of short-term sampling to provide reliable estimates of annual spider assemblages (Araneidae and Thomisidae) present in a Mediterranean site was analyzed, along with the contribution of juvenile data on estimations of spider species richness. A standardized year-long sampling protocol in a one-square-kilometre plot in central Spain yielded reliable Araneidae and Thomisidae inventories. To compare sampling design efficiencies, the degree of completeness of collected annual inventories was estimated, along with an "optimal sampling" selection of months, i.e, the minimum number of months indicating most accurately the number of species present throughout the year. The completeness of spring-month sampling, as well as that of every month, was also estimated. Calculations both included and excluded immature stages. When multiple localities must be sampled and field work minimized, a one-month spring sampling protocol reasonably estimates the entire spring fauna, allowing effective comparisons between sites during the richest period. Our results indicate that juveniles must be included in the sample in order to obtain reliable estimates of species richness, and they should be stored apart from adults in order to analyze them separately as advances in their identification are achieved.

**Keywords:** species richness inventory, Araneae, Araneidae, Thomisidae, short-term sampling, juveniles, species richness estimations

## INTRODUCTION

Although diversity patterns across taxa do not necessarily correlate (Reid, 1998; French, 1999; Kotze & Samways, 1999; Sætersdal *et al*., 2003), management and design of biodiversity conservation strategies are frequently based on information derived from some well known taxa. It follows, then, that conservation policy in general could be enhanced by improving current knowledge of spatial biodiversity patterns of those taxa, such as Arthropods, which account for the greatest part of biodiversity (Kremen *et al*., 1993); consequently, much more corroborative field survey work must be still done (Koch *et al*., 2000). However, reliable field sampling of such highly diverse taxa is no simple task. A long-term, intensive inventory, involving many sample sites in an extensive territory is often impossible for hyperdiverse groups, due to resource (mainly time and money) limitations. Hence, short-term sampling programs capable of reliably identifying all species present in a site are needed. To develop such programs, the seasonal dynamics of the studied assemblage must be well-known (e. g. Landau *et al*., 1999; Moreno & Halffter 2000; Cardoso, 2004).

Although for some taxa, short-term samplings have proved to be useful (e. g. ground beetles: Maelfait & Desender, 1990; moths: Landau *et al*., 1999), some authors have argued that a reduction in spider sampling effort should not imply a decrease in the seasonal width of the sampling protocol (Churchill & Arthur 1999; Riecken, 1999). However, in Mediterranean areas, where summer heat and drought determine a bimodal arthropod species richness distribution, the species richness peaks in spring, and is

followed by a smaller autumn peak that does not seem to add many new species (Shapiro, 1975; Urones & Puerto, 1988; Molina, 1989; Cardoso, 2004). Thus, the gathering of faunistic information from a variety of Mediterranean sites most quickly and efficiently would require the examination of the reliability of short-term sampling design estimations of total species numbers and their comparison with that which could otherwise be obtained over the course of an entire year. We did this by studying the Araneidae and Thomisidae assemblages in a Mediterranean site over a complete annual life cycle.

Another important feature of arthropod sampling, especially when dealing with spiders, is the treatment of immature stages. Usually juveniles are discarded in spider biodiversity studies (e. g. Jerardino *et al*., 1991; Toti *et al*., 2000; Sørensen *et al*., 2002) because they are difficult to identify (Coddington *et al*., 1996; Dobyns, 1997). However, some authors have kept undeveloped stages in the laboratory until maturity in order to include their number in their analysis (e. g. Urones & Puerto, 1988). As juvenile numbers may profoundly influence temporal and spatial spider biodiversity patterns, one must be careful when juveniles are used to compare assemblages and it has been suggested that they should be analyzed separately (Norris, 1999). However, the inclusion of juveniles seems to be necessary in order to obtain reliable short-term sampling estimates of whole-year species richness and composition (Toti *et al*., 2000; Scharf *et al*., 2003). Thus, as Coddington *et al*. (1996) recommended, we have also examined the effects of including juveniles on species richness estimations, in two families of an abundant non specialist predator functional guild (Wise, 1993) in which the identification of juveniles is feasible (i.e. Araneidae and Thomisidae).

The aims of this study are: i) to analyse the capacity of several short-term sampling designs to provide reliable estimates of spiders diversity in a Mediterranean area and ii) to study the effect of juveniles on species richness estimations.

## METHODS

**Study site. —** The study was carried out from April 2003 to March 2004 in a 1 $km^2$ site in central Spain, in the southeast of the Comunidad de Madrid (Perales de Tajuña, 40º14´25´´N 3º23´38´´W). The vegetation is at present dominated by kermes-oaks (*Quercus coccifera*), with a dense shrub undergrowth of *Rosmarinus officinalis* and *Stipa tenacissima*. This sampling site is at 600 meters elevation, with a Mediterranean climate and limestone substratum.

**Sampling protocol. —** Two spider families, Araneidae and Thomisidae, have been studied. These families were selected due to the ease of identification of their juveniles (see below), and because accurate inventories could be obtained. Jiménez-Valverde & Lobo (2005) demonstrate that reliable inventories of these two taxonomic groups can be gathered in one-square kilometer Mediterranean sampling sites. Briefly, the 1 $km^2$ sampling plot was divided into 2500 subplots of 400 $m^2$; 20 of these subplots were chosen at random, and a subsample unit carried out in each. A subsample unit was defined as: i) a one-person sweep of the herbaceous vegetation and shrub during 15 minutes, ii) a one-person beating of bushes and small trees and branches during 15 minutes, and iii) the running, during 48 hours, of 4 pitfall traps 11.5 cm wide and 1 litter in volume, each separated by 10 m from the others. Traps were filled with water, and a few drops of detergent added to break the surface tension so as to prevent the spiders from escaping. Sampling was always done by the same person (A. J.-V.) in order to

avoid possible differences due to the effect of the collector (Norris, 1999). Rainy and windy days were avoided in order to prevent a reduction in the efficiency of the sampling methods. This protocol gathers reliable Araneidae and Thomisidae inventories and is highly repeatable because the data gathered are related to sampling effort measure (see details in Jiménez-Valverde & Lobo, 2005). This protocol was performed once a month (except in April, when it was done twice) in order to study the seasonal variation of the richness and composition of the assemblage (dates of sampling: Ap-1: 1/IV/03-9/IV/03, Ap-2: 22/IV/03-29/IV/03, Ma-Ju: 31/V/03-6/VI/03, July: 16/VII/03-23/VII/03, Aug: 13/VIII/03-26/VIII/03, Sep: 17/IX/03-26/IX/03, Oct-Nov: 22/X/03-14/XI/03, Dec: 2/XII/03-12/XII/03, Jan: 9/I/04-21/I/04, Feb: 11/II/04-1/III/04, Mar: 10/III/04-22/III/04). In total, 220 subsample units were carried out.

**Juvenile sampling. —** Juveniles that could be identified to the species level were included in the analysis. This was possible for many araneid and some thomisid species which have a distinguishing, characteristic color pattern, and for some genera represented by only one species in the Iberian Peninsula (i.e. *Mangora acalypha* (Walckenaer, 1802), *Runcinia grammica* (Koch, C.L., 1837) and *Synaema globosum* (Fabricius, 1775)).

**Data analysis. —** The degree of completeness of the collected annual inventory was estimated, as well as an "optimal sampling" selection of months, a minimum, to identify the year-round number of species. The completeness of the sample gathered during the spring months, as well as that of every month, was also estimated by means of i) species accumulation curves, ii) non-parametric estimators, and iii) the estimated area under the truncated lognormal species-abundance distribution curve.

Species accumulation curves were built (see Gotelli & Colwell, 2001) using three sampling effort surrogates, which are the number of individuals (N), the number

subsamples (S), and the number of months (M; used only for the yearly sampling). However, the accuracy of monthly samplings was evaluated by using only the number of subsamples units (i.e. the number of subplots, 20 every month). The order in which sampling effort units were added was randomized 500 times to build smoothed curves using the EstimateS 5.0.1 software (Colwell, 1997). The asymptotic value of the accumulation curves obtained was estimated using both the Clench and Weibull equations (Soberón & Llorente, 1993; Colwell & Coddington, 1994; Flather, 1996; León-Cortés *et al*., 1998; Peterson & Slade, 1998). These models were fitted to the data through non-linear regression using the Simplex & Quasi-Newton algorithm (StatSoft, 2001).

Four non-parametric species richness estimators were calculated: the first- and second-order jackknife (Jack1 and Jack2), the abundance-based coverage (ACE) and the incidence-based coverage estimator (ICE). These four estimators have performed relatively well in numerous studies (Palmer, 1990; Palmer, 1991; Coddington *et al*., 1996; Boulinier *et al*., 1998; Toti *et al*., 2000; Walther & Martin, 2001; Borges & Brown, 2003; Brose *et al.*, 2003; Chiarucci *et al.*, 2003). Detailed descriptions of the estimators can be found in Colwell & Coddington (1994) and Colwell (1997).

The truncated lognormal distribution model was fitted to the abundance data (Magurran, 1988) in order to estimate the area under the curve, or the total number of species that could be expected if an exhaustive collection effort was carried out (Fagan & Kareiva, 1997). Octaves were defined as $\log_2$ following Preston (1948, 1962) (see Lobo & Favila, 1999) and the Kolmogorov-Smirnov one-sample test with Lilliefors corrected critical values was used to compare the observed and expected patterns of species abundance distributions (Tokeshi, 1993).

In order to cluster monthly inventories according to their taxonomic resemblance, the Sørensen similarity coefficient was used, taking into account presence/absence data. Because true absences of species are difficult to verify in inventories, this coefficient was selected because it doubles the weight of double presences (Legendre & Legendre, 1998). The Bray-Curtis coefficient (quantitative version of the Sørensen coefficient, Legendre & Legendre, 1998) was also used when considering abundance data. Because both cluster analyses generate similar dendrograms, only the results from the Bray-Curtis coefficient are shown. Ward's method was used as linkage rule, a method which tries to minimize the difference between the sum of the squared distances of cases and the mean values of the clusters to which they are assigned (Legendre & Legendre, 1998). We used NTSYSpc 2.11 (Rohlf, 2000) and STATISTICA (StatSoft, 2001) software in these analyses.

## RESULTS

**Faunistic composition. —** A total of 1599 individuals were captured, 1471 of them juveniles (92%); of these, 1241 (84%) were used in the analysis, as the rest (230) were impossible to identify unambiguously to species level. In all, 20 species were collected; three of them captured only as immature stages (*Aculepira armida* (Audouin, 1826), *Gibbaranea* sp. Archer, 1951 and *Hypsosinga albovittata* (Westring, 1851)).

**Seasonal variation of the assemblage. —** The seasonal number of both species and individuals from adult+juvenile data peaked in spring (Ap-1, Ap-2 and Ma-Ju). Numbers decreased in summer (Jul, Aug and Sep), while a nearly constant recovery occurs afterwards (Fig. 1A). Interestingly, the number of species was quite stable from March to May-June, unlike the number of individuals. On the other hand, in the case of

adults-only data, a clearly defined peak both in the number of species and individuals occurred in late spring (Ap-2 and Ma-Ju), followed by a sharp decline that persists throughout the rest of the year (Fig. 1B). In the case of adults-only data, the seasonal variation in the number of species was more erratic, as there were months in which no adult spider was collected, as in October-November or February.

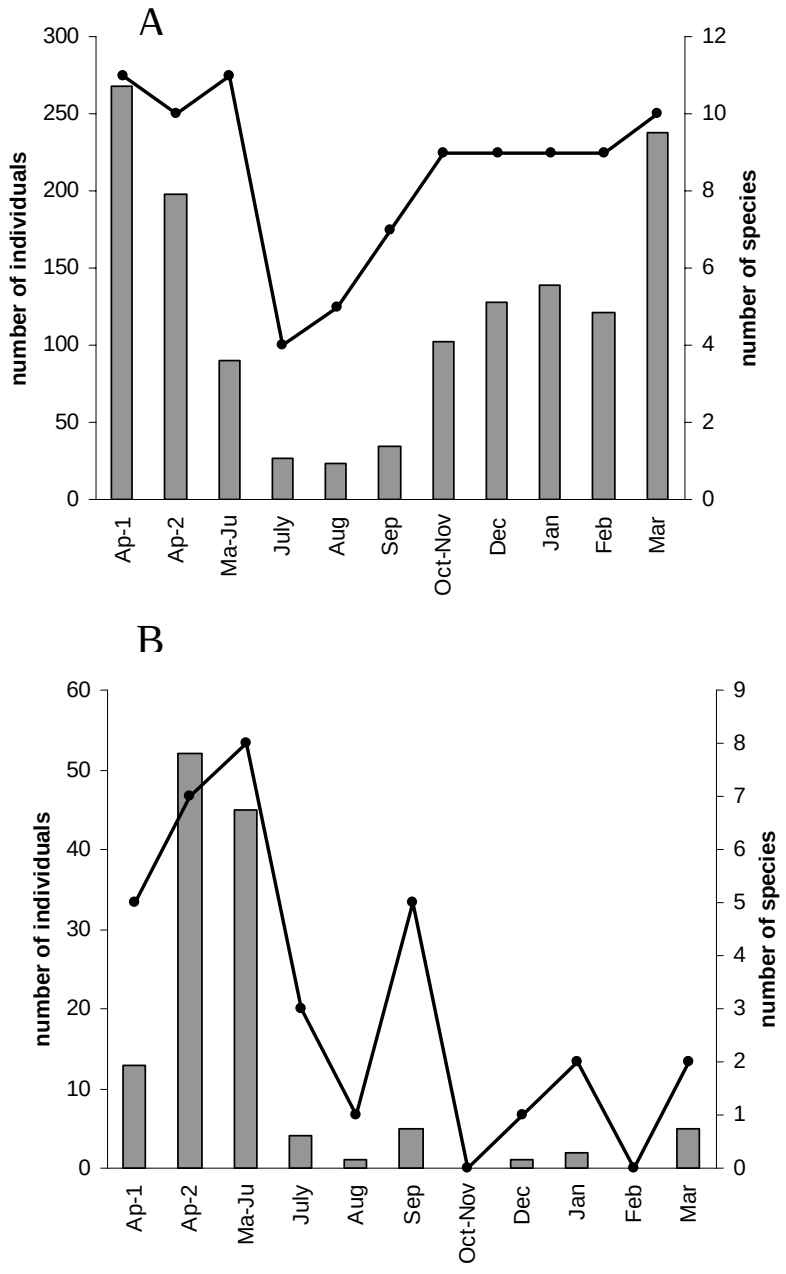



Figure 1.- Seasonal variation of the number of species (lines) and number of individuals (columns) of the spider assemblage studied including juveniles (A) and adults-only specimens (B).

Months grouped according to faunistic similarity, and which include juveniles, belonged to either a group that contains the less-species-rich summer inventories, with fewer individuals (Jul, Aug and Sep), or to another with the remaining inventories (Fig. 2A), in which the composition of the fauna identified in the species-rich months was very similar. Adults-only data generated a similar tree diagram, in which species-rich spring inventories are again clearly associated (Fig. 2B).

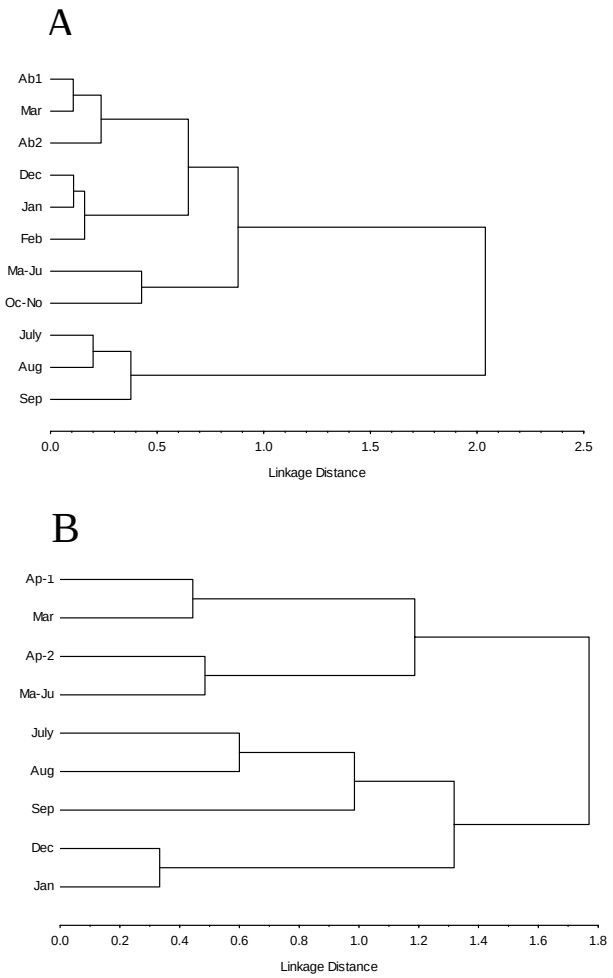



Figure 2.- Dendrograms showing the similarity of monthly samplings using the Bray-Curtis similarity coefficient as resemblance measure and Ward´s method as linkage rule (A: juveniles included; B: juveniles excluded). Note that no adults were captured neither in Oct-Nov nor in Feb, so such samples are excluded from dendrogram B.

**Accuracy of the annual sampling**. **—** Lognormal estimations, accumulation curve functions and non-parametric estimators all showed that observed richness represented around 80% of that estimated when both adult and juvenile data were considered (Table 1). Accumulation curves nearly reached an asymptote (Fig. 3); the estimates by the various methods suggest that only about three to five more species would be collected if an exhaustive sampling effort was carried out. Adults-only data lead to higher estimation scores (Mann-Whitney U test = 2.63, $p = 0.008$) and also to a greater range of variation (26.5 ± 1.2 species; mean ± SE) than those obtained from all-specimens data (22.7 ± 0.3 species). This adults-only data leads to an observed species richness of around 60% of estimated species richness, and would indicate that almost 10 more species should have to be added to the inventory (Table 1). However, a clear pattern in the variation in the level of completeness, according to the measure of sampling effort used (subsamples, months or individuals; Table 1), was not apparent. Adult+juvenile species abundance data closely fits a lognormal truncated distribution ($D$=0.051, ns), as well as the species abundance distribution of adults ($D$=0.063, ns), indicating a very similar total number of species (24 or 25 species; see Table 1).

Non-parametric estimators with adults-only data tended to overestimate species richness and lead to considerably different predictions of species richness. As adults-only data produced a greater proportion of rare species (singletons and doubletons) (Table 1, Fig. 3), the inventories so derived seemed, *a priori*, to be more incomplete than those derived from all-specimens data. Moreover, Clench and Weibull estimations were affected by the difference in the shapes of the accumulation curves (see Fig. 3). A high percentage of singletons in the data resulted in an extremely gradual rate of species addition, leading to steeper accumulation curves and greater function slopes at the end of the curves. The Clench adults-only final slopes range from 0.7 to 0.03; all-specimen

Table 1.- Sampling design results, with and without juveniles (A+J, A) and calculated for various sampling effort units (subsamples, months and individuals): number of observations or sampling units, number of species observed (*Sobs*), number and percentage over *Sobs* of singletons (species with only one individual) and doubletons (species with only two individuals), number of species predicted (*Spred*) and percentage over *Sobs* of the Clench and Weibull functions, number of species predicted and percentage over *Sobs* of the ICE (incidence-based coverage), ACE (abundance base coverage), Jack1 (first-order Jackknife) and Jack2 (second-order Jackknife) nonparametric estimators, and number of species predicted and percentage over *Sobs* by the lognormal abundance distribution.

| | **Complete sampling** | | | | | | **Optimal sampling** | | | | **Spring sampling** | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| | A+J Subsamples | A+J Months | A Subsamples | A Months | A+J Individuals | A Individuals | A+J Subsamples | A Subsamples | A+J Individuals | A Individuals | A+J Subsamples | A Subsamples | A+J Individuals | A Individuals |
| Number of observations | 220 | 11 | 220 | 11 | 1369 | 128 | 80 | 100 | 514 | 68 | 80 | 80 | 794 | 115 |
| *Sobs* | 20 | 20 | 17 | 17 | 20 | 17 | 20 | 17 | 20 | 17 | 15 | 10 | 15 | 10 |
| Singletons | 3 | 3 | 7 | 7 | 3 | 7 | 7 | 10 | 7 | 10 | 1 | 2 | 1 | 2 |
| % | 15 | 15 | 41.2 | 41.2 | 15.0 | 41.2 | 35.0 | 58.8 | 35.0 | 58.8 | 6.7 | 20.0 | 6.7 | 20.0 |
| Doubletons | 3 | 3 | 3 | 3 | 3 | 3 | 3 | 1 | 3 | 1 | 1 | 1 | 1 | 1 |
| % | 15 | 15 | 17.6 | 17.6 | 15.0 | 17.6 | 15.0 | 5.9 | 15.0 | 5.9 | 6.7 | 10.0 | 6.7 | 10.0 |
| *Spred* Clench fit | 21.2 | 23.5 | 22.2 | 31.9 | 20.9 | 21.6 | 22.6 | 23.4 | 22.0 | 23.3 | 16.0 | 11.3 | 15.6 | 10.9 |
| % | 94.2 | 85.2 | 76.4 | 53.3 | 95.7 | 78.8 | 88.6 | 72.7 | 90.9 | 72.8 | 93.9 | 88.8 | 95.8 | 91.3 |
| *Spred* Weibull fit | 24.3 | 22.6 | 29.1 | 22.1 | 24.8 | 39.3 | 33.3 | 46.0 | 54.4 | 29.0 | 15.9 | 10.2 | 16.2 | 10.4 |
| % | 82.4 | 88.4 | 58.5 | 77.1 | 80.5 | 43.2 | 60.0 | 36.9 | 36.8 | 58.7 | 94.5 | 97.9 | 92.5 | 96.5 |
| ICE | 22.4 | 22.1 | 27.4 | 23.2 | - | - | 31.1 | 35.3 | - | - | 15.4 | 11.2 | - | - |
| % | 89.4 | 90.5 | 61.9 | 73.1 | - | - | 64.2 | 48.1 | - | - | 97.3 | 89.2 | - | - |
| ACE | 22.3 | 22.3 | 28.6 | 28.6 | - | - | 34.5 | 34.8 | - | - | 15.4 | 11.3 | - | - |
| % | 89.8 | 89.8 | 59.4 | 59.4 | - | - | 57.9 | 48.8 | - | - | 97.1 | 88.5 | - | - |
| Jack1 | 23.0 | 22.7 | 24.0 | 23.4 | - | - | 26.9 | 26.9 | - | - | 16.0 | 12.0 | - | - |
| % | 87.0 | 88.0 | 70.9 | 72.8 | - | - | 74.3 | 63.2 | - | - | 93.8 | 83.5 | - | - |
| Jack2 | 23.0 | 20.8 | 27.9 | 23.9 | - | - | 30.8 | 35.7 | - | - | 16.0 | 13.0 | - | - |
| % | 87.0 | 96.3 | 60.8 | 71.0 | - | - | 64.8 | 47.6 | - | - | 93.7 | 77.2 | - | - |
| Lognormal estimation | - | - | - | - | 24.4 | 25.0 | - | - | 37.9 | 27 | - | - | 16.2 | 10.5 |
| % | - | - | - | - | 82.0 | 68.1 | - | - | 52.8 | 63.0 | - | - | 92.6 | 95.2 |

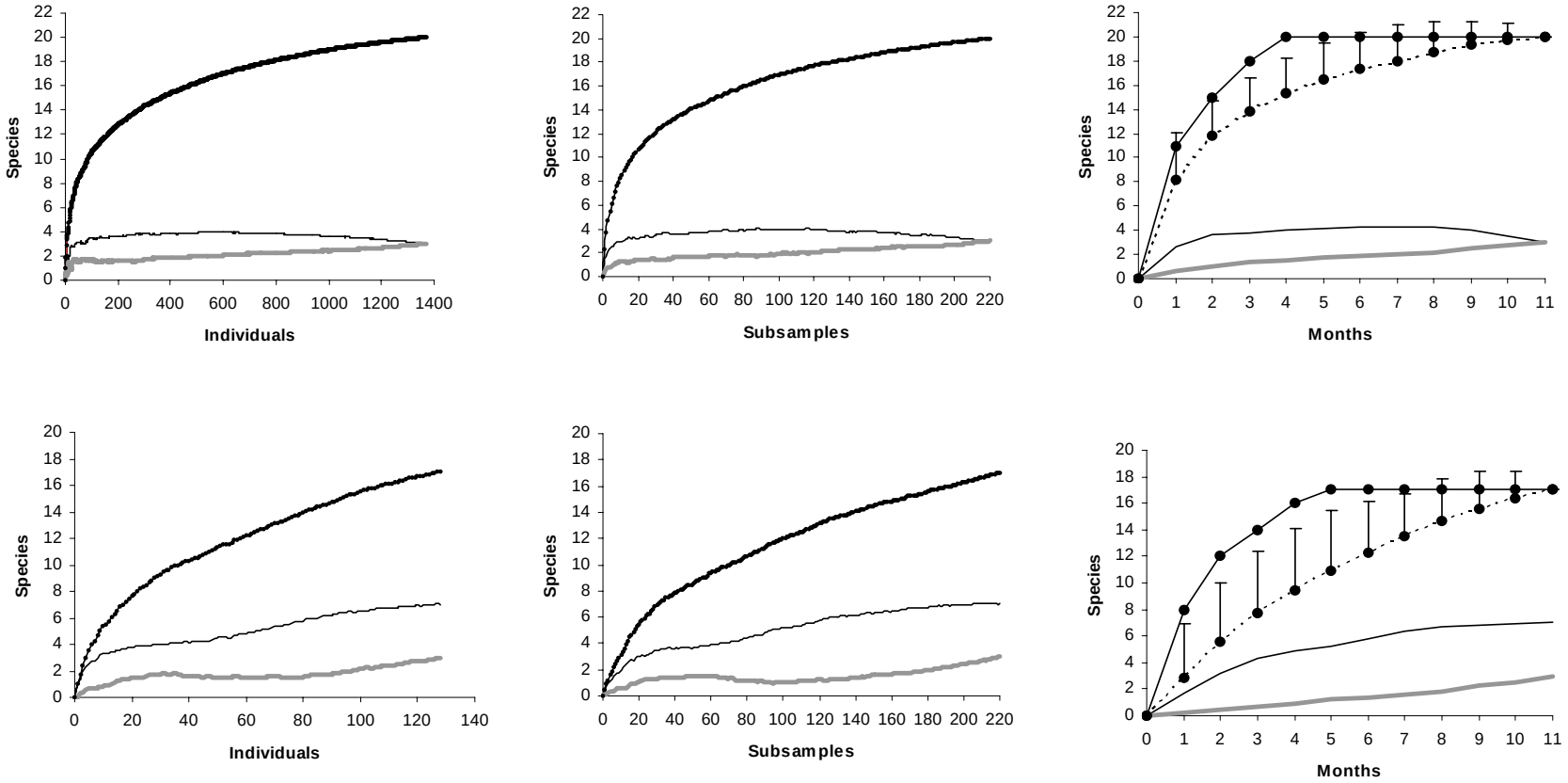



Figure 3.- Randomized accumulation curves (500 times) for the annual inventory employing different sampling effort measures (A: juveniles included; B: juveniles excluded; thick line: species observed; thin line: number of singletons; grey line: number of doubletons). For the curve using the number of months as sampling effort measure the +95% confidence intervals have been indicated for the randomized curve, as well as the curve produced by ordering months according to a complementary criterion ("optimal sampling", solid line).

final slopes from 0.2 to 0.001; Weibull adults-only final slopes range from 0.6 to 0.02; all-specimen final slopes from 0.3 to 0.002. Thus, adults-only data clearly lead to overestimations of species richness, especially in the case of the Weibull model (Table 1).

Clench and Weibull models extrapolated to values of even twice the number of individuals actually collected yield curves that were still non-asymptotic (especially the latter; Fig. 4). This indicates that doubling the sampling effort would not make the collected inventories appreciably more representative. The "reasonably true" species richness was therefore chosen as 24, for purposes of comparison.

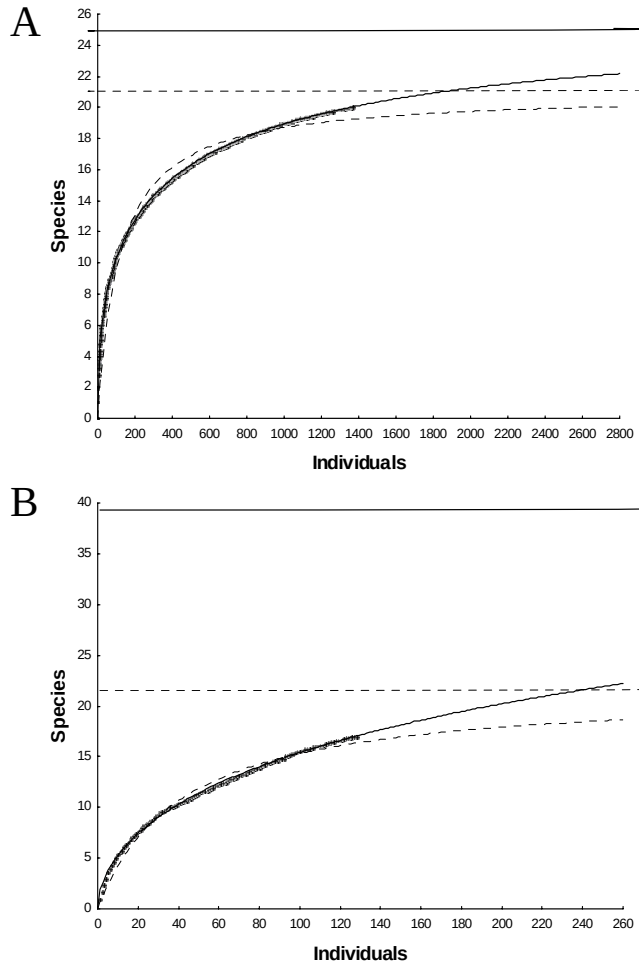



Figure 4.- Extrapolations of the Clench and Weibull models by doubling the annual number of individuals actually collected (dotted curve: Clench model; solid curve: Weibull model; dotted horizontal line: Clench´s asymptote; solid horizontal line: Weibull´s asymptote). A: considering juveniles; B: excluding juveniles.

**Accuracy of the optimal sampling**. **—** By an iterative procedure, months were selected sequentially, according to their contribution to the total accumulated species richness, until no more new species were added (Table 2). A seasonal optimal sampling selection indicates that, in accordance with all-specimen data, four months is the minimum period necessary to guarantee the capture of a number of species equal to that collected in the year-long sampling. If juveniles are not included, five months are necessary. Three sampling periods (Ap-1, Ma-Ju and Sep) were found to be common to the results from both data sets. A random selection (500 randomizations) of the same number of months produced a significantly smaller number of observed species (Fig. 3). Estimates derived from this optimal sampling period which include juvenile data were significantly higher than those from year-long periods (Mann-Whitney U test = 19.5, $p$= 0.004) but not from adults-only data which excluded juveniles (Mann-Whitney U test = 41.0, $p$= 0.1); the ranges of variation in the optimal sampling period are also higher (adults+juveniles 32.6 ± 3.2, adults 31.3 ± 2.5). This means that there are 13 or 14 more species estimated than observed (Table 1). Thus, data from an optimal month selection operated on by predictive methods produces higher and more erratic species richness estimations than reasonably true scores. Moreover, accumulation curves did not approach an asymptote, especially when derived from adults-only data. Singletons and doubletons accounted for a higher proportion of species numbers than in the annual sampling (Table 1); the steep slope and proportion of singletons in adults-only data is especially remarkable, leading to some strikingly disproportionate estimates in the cases of the Weibull function and non-parametric estimators. Although species-abundance data closely fitted a lognormal distribution (adults+juveniles: $D$=0.09, ns; adults: $D$=0.10, ns), this method also overestimated the year-round number of species. Thus, the estimations generated with the inventory obtained from an optimal sampling

procedure such as this generally produce overestimations of the true year-round species richness. Only the estimates generated by the Clench model, using both subsamples and individuals, yielded reasonable values (see Table 1).

Table 2.- Results of an iterative complementarity procedure in which the inventories of each month were sequentially selected according to its contribution to the species richness.

| **Adults + juveniles** | | | |
|---|---|---|---|
| **Iteration** | **Month** | **Species** | **Accumulated species** |
| 1 | April 1 | 11 | 11 |
| 2 | September | 4 | 15 |
| 3 | February | 3 | 18 |
| 4 | May-June | 2 | 20 |

| **Adults** | | | |
|---|---|---|---|
| **Iteration** | **Month** | **Species** | **Accumulated species** |
| 1 | May-June | 8 | 8 |
| 2 | September | 4 | 12 |
| 3 | April 1 | 2 | 14 |
| 4 | July | 2 | 16 |
| 5 | December | 1 | 17 |

**Spring sampling**. **—** According to both species richness and compositional variation (Figs. 1 & 2) spring sampling was taken to be the faunistically similar period within which the annual species richness peaked (Ap-1 + Ap-2 + Ma-Ju + Mar). Accumulation curve functions, non-parametric and lognormal estimates (adults+juveniles: $D$=0.06, ns; adults: $D$=0.10, ns) all indicated that observed richness represented around 95% of that estimated from all-specimen data (Table 1). Estimates from adults-only data were slightly less stable, falling below the observed all-specimen richness. Clearly, excluding juveniles leads to an underestimation of spring species richness, which, at 16, would be considered to be its reasonably true value. However, this low value represents, respectively, only 80% and 67% of the observed and estimated annual species richness.

**Monthly samplings. —** Adults-only data lead to underestimates of species in some months (Table 3). Moreover, there were some months in which only juveniles were collected (October-November and February). On average, five species (around 50%) were absent in spring months, juveniles excluded, while almost six species must be so considered (77% of species richness) in the remaining months. Thus, because excluding juveniles leads to underestimates of species richness values, only data that included juveniles were analyzed.

In general, monthly samplings were accurate, judging from a completeness of monthly inventories that varied around a value of 80% (Table 3, Fig. 5). The exceptions were inventories characterized by a high proportion of singletons from three months (August, February and, specially, September), which produced much steeper accumulation curves than those found for the other months (Table 3). The negative correlation of mean completeness with the standard deviation of estimates (*Spearman rank coefficient* = -0.80, $p < 0.01$) indicates that the more incomplete an inventory, the poorer the concordance (and so, accuracy) of estimators.

A single-month sampling in spring captured around 70% of the observed spring fauna, and around 50% of the year-round fauna (Table 3). Observed spring species richness varied between 62% and 69% of the estimated reasonable true spring spider species richness (16 species), and between 42% and 46% of the yearly estimated richness (24 species). Moreover, monthly spring estimations varied from 62%-100% of the estimated true spring species richness, and from 42%-67% of the estimated true yearly species richness. However, estimations from both months and estimators varied greatly in their reliability; March was the least reliable of spring months, while the Jack2 estimator tends to produce an estimation extremely similar to the reasonably true spring species richness score (16 species, see above) (Table 3).

Table 3.- Results for the monthly samplings, with and without juveniles (because not including immature stages leads to a very low sample size, no estimator has been calculated in this case, see text): number of individuals collected, number of species observed (*Sobs*), number and percentage over *Sobs* of singletons and doubletons, number of species predicted (*Spred*) and percentage over *Sobs* of the Clench and Weibull functions, number of species predicted and percentage over *Sobs* of the ICE, ACE, Jack1 and Jack2 nonparametric estimators, and percentage of species observed over the total observed and predicted for spring and the annual sampling.

| **Adults+juvenils** | **April 1** | **April 2** | **May-June** | **July** | **August** | **September** | **October-November** | **December** | **January** | **February** | **March** |
|---|---|---|---|---|---|---|---|---|---|---|---|
| Number of Individuals | 268 | 198 | 90 | 27 | 23 | 35 | 102 | 128 | 139 | 121 | 238 |
| *Sobs* | 11 | 10 | 11 | 4 | 5 | 7 | 9 | 9 | 9 | 9 | 10 |
| Singletones | 3 (27.3%) | 1 (10.0%) | 3 (27.3%) | 1 (25.0%) | 3 (60.0%) | 5 (71.4%) | 2 (22.2%) | 2 (22.2%) | 2 (22.2%) | 5 (55.6) | 2 (20.0%) |
| Doubletones | 1 (9.1%) | 0 | 0 | 1 (25.0%) | 1 (20.0%) | 0 | 2 (22.2%) | 1 (11.1%) | 0 | 0 | 1 (10.0%) |
| *Spred* Clench fit | 12 (91.7%) | 11 (90.9%) | 12 (91.7%) | 5.1 (78.4) | 7.94 (63.0%) | 11 (63.6%) | 12 (75.0%) | 11 (81.8%) | 10 (90.0%) | 12 (75.0%) | 12 (83.3%) |
| *Spred* Weibull fit | 14 (78.6%) | 13 (76.9%) | 12 (91.7%) | 4.66 (85.8%) | 11.24 (44.5%) | 253 (2.8%) | 11 (81.8%) | 10 (90.0%) | 9 (100%) | 230 (3.9%) | 13 (76.9) |
| ICE | 14 (78.6%) | 13 (76.9%) | 13 (84.6%) | 5.09 (78.6%) | 12.03 (41.6%) | 30 (23.3%) | 11 (81.8%) | 10 (90.0%) | 11 (81.8%) | 19 (47.4%) | 11 (90.9%) |
| ACE | 15 (73.3%) | 10 (100%) | 13 (84.6%) | 4.69 (85.3%) | 11 (45.4%) | 31 (22.6%) | 11 (81.8%) | 11 (81.8%) | 10 (90.0%) | 20 (45.0%) | 12 (83.3%) |
| Jack1 | 14 (78.6%) | 13 (76.9%) | 14 (78.6%) | 4.95 (80.8%) | 7.85 (63.7%) | 12 (58.3%) | 11 (81.8%) | 12 (75.0%) | 12 (75.0%) | 14 (64.3%) | 12 (83.3%) |
| Jack2 | 16 (68.7%) | 16 (62.5%) | 16 (68.7%) | 5 (80.0%) | 9.7 (51.5%) | 16 (43.7%) | 10 (90.0%) | 12 (75.0%) | 12 (75.0%) | 18 (50.0%) | 12 (83.3%) |
| *S* observed spring % | 73.3 | 66.7 | 73.3 | 26.7 | 33.3 | 46.7 | 60.0 | 60.0 | 60.0 | 60.0 | 66.7 |
| *S* observed total % | 55.0 | 50.0 | 55.0 | 20.0 | 25.0 | 35.0 | 45.0 | 45.0 | 45.0 | 45.0 | 50.0 |
| *S* predicted spring % | 68.8 | 62.5 | 68.8 | 25.0 | 31.3 | 43.8 | 56.3 | 56.3 | 56.3 | 56.3 | 62.5 |
| *S* predicted total % | 45.8 | 41.7 | 45.8 | 16.7 | 20.8 | 29.2 | 37.5 | 37.5 | 37.5 | 37.5 | 41.7 |
| **Adults** | | | | | | | | | | | |
| Number of Individuals | 13 | 52 | 45 | 4 | 1 | 5 | 0 | 1 | 2 | 0 | 5 |
| *Sobs* | 5 | 7 | 8 | 3 | 1 | 5 | 0 | 1 | 2 | 0 | 2 |
| Singletones | 3 (60.0%) | 2 (28.6%) | 3 (37.5%) | 2 (66.7%) | 1 (100%) | 5 (100%) | - | 1 (100%) | 2 (100%) | - | 0 |
| Doubletones | 1 (20.0%) | 0 | 0 | 1 (33.3%) | - | 0 | - | - | 0 | - | 1 (50%) |

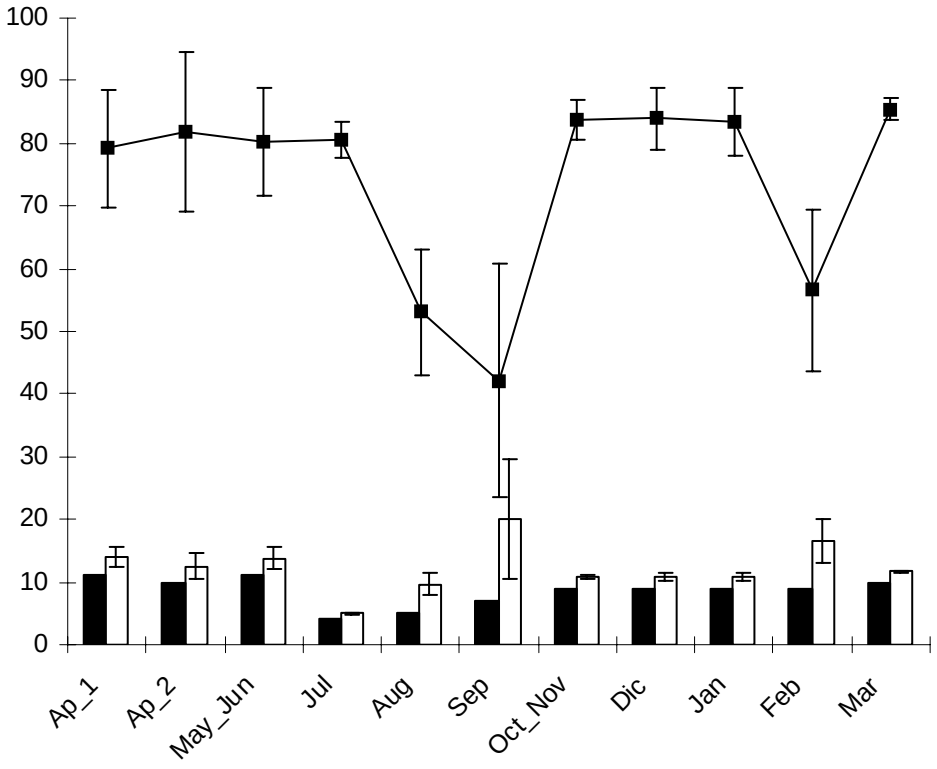



Figure 5.- Number of observed species (black columns), mean of estimations ± SD (white columns) and mean percentage of completeness ± SD (square dots) over the annual total of species. All calculations made ignoring Weibull scores because of extremely high values (see Table 3).

**Similarity among sampling protocols**. **—** A gradient in Bray Curtis faunistic dissimilarity is observed from spring sampling, the more similar to the year-long sampling, to non-spring months (Fig. 6). The gradient in species richness similarity shows the same pattern with the exception of the optimal sampling which, obviously, marks the same species richness as the year-long sampling.

**The effect of including juveniles**. **—** For purposes of comparison, the accumulated number of species using adults-only and adults+juveniles data were plotted together against number of subsamples (Fig. 7A). Not including juveniles produced lower observed species richness scores at every stage of the sampling process, increasing the difference as long as the sampling effort raised until an inflexion point

(~70 subsamples) after which the advantage of including juveniles progressively disappeared (Fig 7A). But, is the favorable effect of including juveniles attributable to the increase in sample size or to an intrinsical property of juveniles? We resampled 20 times the 1369 annual individuals (adults and juveniles) at *n*=128 (the number of annual adults) calculating the number of observed species and estimating Clench predictions in order to detect the possible effect of sample size. The resulting observed scores varied from 9 to 14 (mean number ± SD: 10.9 ± 1.4), while the estimations varied from 9.7 to 17.5 (13.2 ± 2.1). These values are significantly lower than those using the 128 adults as well as the 1369 adults and juveniles (see Table 1). A comparative examination of adult and adult-juvenile individual-based accumulation curves clearly shows the same pattern; i.e. that the same number of individuals yielded lower observed and predicted species richness scores using juveniles data than when only adults are used (Fig. 7B).

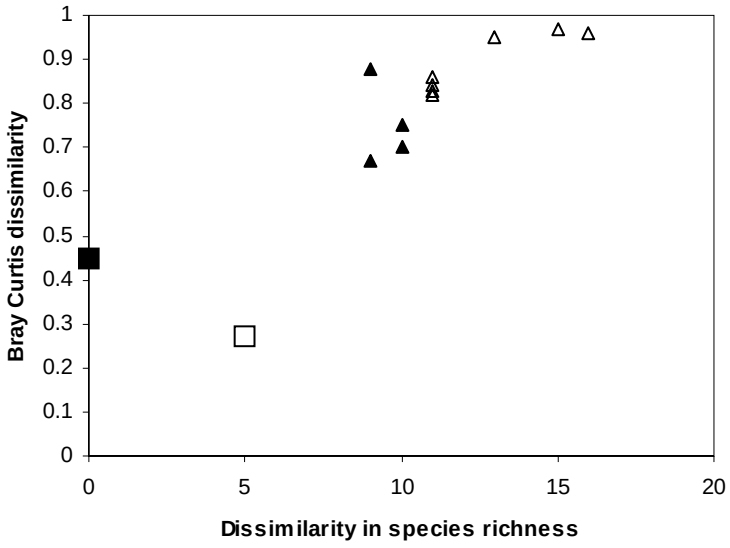



Figure 6.- Difference between the number of annual collected species (20 species) and the collected richness in different inventories (i.e. dissimilarity in richness scores; X axis), and Bray Curtis faunistic dissimilarity distance (Y axis) of the different sampling protocols with respect to the year-long one (black square: optimal sampling, white square: spring sampling, black triangles: spring sampling months, white triangles: rest of months).

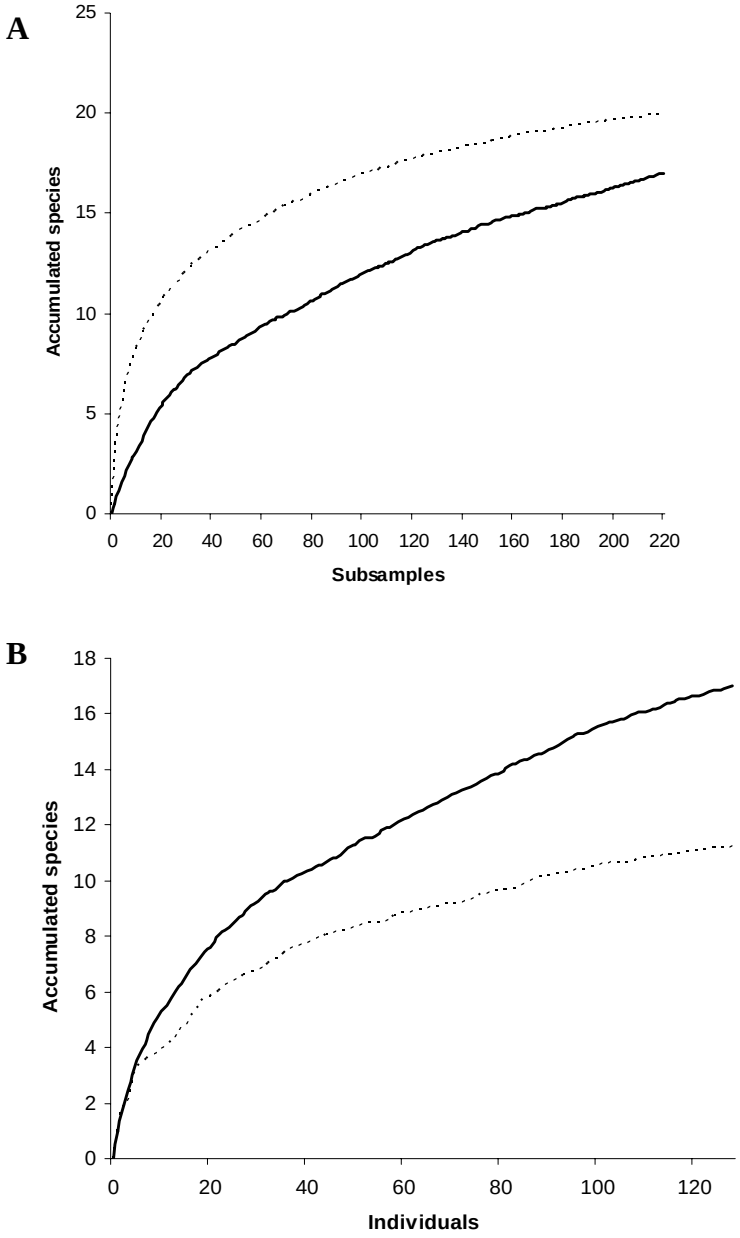



Figure 7.- Randomized (500 times) subsample-based (A) and individual-based (B) accumulation curves for the annual inventory (solid line: juveniles excluded; dotted line: juveniles included).

## DISCUSSION

The year-round survey carried out for this study yielded a good representation of the spider fauna of this kermes-oak forest. The seasonal pattern observed is typical of Mediterranean habitats where summer ushers in adversely high temperatures and

pronounced drought (Shapiro, 1975; Abraham, 1983). In such conditions species richness peaks twice. One peak occurs in spring, when a maximum number of species is attained. Many juveniles appear at the beginning of this season, but only a small fraction reach maturity in May-June, when the adult peak of species richness is reached. In July there is a considerable decline in species and individual numbers, although later some new species appear in very small numbers during August and September. This population fluctuation is reflected in the similarity of spring and year-round assemblage structure (Fig. 6). Juveniles of the spring species appear again in autumn. Indeed, cluster analysis of data that included juveniles lead to a clear association of all monthly inventories except during July, August and September. However, adults-only data produced a separate cluster in winter months (December and January). The existence of these two phenologic peaks in species richness enhances the opportunities to find an effective short-term sampling design in order to reach a trade-off between reliability of inventories and survey costs.

Unfortunately, long-term intensive sampling may not be affordable in biodiversity surveys, especially those with multiple sampling points. A shorter-length optimal sampling protocol that collected the same number of species as year-round sampling would reduce the sampling effort. The main problem with this strategy is that it significantly increases the proportion of singletons and the steepness of the accumulation curves, generating a sample with a species-abundance relationship different to that of the observed in the year-round assemblage (Fig. 6), biasing the estimations. Moreover, with many sampling locations, such an optimal protocol is still difficult to carry out.

Another strategy that would similarly reduce the required field effort, but that reduces the seasonal coverage of samplings, would be the limiting of surveys to the

richest season, such as spring. In the case of this study, spring sampling identified this season´s fauna quite well, as indicated by the low range of variation of all estimators used, and the high degree of similarity between observed and predicted richness scores. However, these estimators do not allow the further extrapolation of the sampled universe. Thus, although a reliable figure for the entire spring fauna can be obtained, the year-round species richness remains unknown.

Monthly estimates seem to be dependent on species richness. Predicted and observed species richness differs slightly from one to another of the richest months; this difference is greater in the species-poor summer months. Short-term spring sampling in April or May-June, with only 20 sampling effort units, seems to yield quite good estimations of the observed spring species richness and even of the predicted reasonably true spring species richness. The Jack2 estimator performed well in these extrapolations. Jackknife estimators in general, and Jack2 in particular, have been found to perform quite well, with greater precision, less bias and lesser dependence on sample size than other estimators, by many authors (Palmer, 1990, 1991; Baltanás, 1992; Brose *et al*., 2003; Petersen *et al*., 2003; Chiarucci *et al*., 2003). Accordingly, the use of this estimator is herein suggested to predict spring species richness, after previous assessment of inventory completeness by visual examination of the accumulation curve and a screening of the estimation deviations for the high levels of incompleteness which could alter estimation results. However, if inventories are less complete than the one analyzed herein, other orders of Jackknife estimators should be used (see Brose *et al*., 2003 for a detailed protocol to select estimator algorithms).

Thus, to optimize the field work involved in sampling the spider fauna (Araneidae and Thomisidae) of multiple sites under Mediterranean conditions, an exhaustive sampling protocol in one spring month is herein proposed. This strategy

yields reasonable estimates of the entire spring fauna and, if spatial homogeneity of this observed pattern is preserved sites can be effectively compared, since spring inventories are a good representation both of the annual species richness and faunistic composition. Bearing in mind that a trade-off exists between survey effort and data quality, year-round sampling would be preferable if cost permits. Otherwise, optimal or spring samplings are two viable options; the former to be recommended if purely faunistic and taxonomic information (list of species) is of interest, the latter if a realistic picture of the yearly compositional structure is. If resources are extremely scarce, spring monthly sampling would be the preferred option. Although it is probable that a similar pattern may be common in many spider families, our conclusions must be restricted to Araneidae and Thomisidae because phenological patterns can vary among spider families.

The proportion of immature stages collected in this study, as in other spider studies, is very large (Coddington *et al*., 1996; Kuntner & Baxter, 1997; Cardoso *et al*., 2004; Sørensen, 2004). Because of demographic unavoidable reasons, rare species are more probably to be collected in early stages of the survey when dealing with juveniles, yielding a better representation of the complete assemblage more quickly (Fig. 7A). As the year-round sampling protocol, the effect of include juveniles diminishes the proportion of singletons and increases the asymptotic tendency of the accumulation curve. Moreover, the general shape of both accumulation curves, with and without juveniles, as well as their concomitant changes in steepness are different, because they constitute two different sampling universes (*sensu* Colwell & Coddington, 1994), altering estimates of phenomenological accumulation models (Fagan & Kareiva, 1997; Melo *et al*., 2003). The value of Weibull´s asymptote, a function that fits data quite well (low sum-of-squares and high explained variance; see Flather, 1996 and Jiménez-

Valverde *et al*., 2006), is strongly influenced by the shape of the accumulation curve. Strikingly high estimates are derived from Weibull values obtained from curves whose slopes approach the asymptote more steeply towards their end, as occur in the adults-only accumulation curve. Overfitting is not a desirable property and the Clench model, less flexible and more conservative, approaches the asymptote with greater readiness than does the Weibull function yielding less biased estimations especially for non asymptotic data sets.

Whatever the seasonal coverage of the sampling carried out adult-only specimen numbers greatly reduce the size of sample to be analyzed, and the faithfulness of sample representation of true species richness, while increasing the proportion of rare species (singletons, doubletons). All of the foregoing affects the performance of species richness estimators (Heltshe & Forrester, 1983; Smith & van Belle, 1984; Chao, 1987; Baltanás, 1992; Colwell & Coddington, 1994; Brose *et al*., 2003). Differences observed in the individual-based accumulation curves of adults-only and adults+juveniles data (Fig 7B) are due to differences in the species-abundance relationships of both universes. In the adults+juveniles universe, common species have a higher relative abundance than in the adults-only universe. Thus, the higher observed species richness and more accurate estimations obtained when juveniles are included seem to be mainly due to the unavoidable increase in the number of individuals that thereby results. Definitively considering juveniles is necessary to obtain reliable estimates of species richness except when the sampling effort is so high that formerly unrepresented rare species emerge in the adult-only data set. Estimator robustness with respect to community structure changes is a desirable property (Melo *et al*., 2003) but, until now, no estimator has been shown to be so robust (Keating & Quinn, 1998; Baltanás, 1992). Accordingly, sample size must be large enough to include a high proportion of the true species richness and

to represent the actual assemblage (Baltanás, 1992; Willot, 2001; Brose *et al*., 2003; Melo *et al*., 2003; Petersen & Meier, 2003; Petersen *et al*., 2003); and this can be better achieved including juveniles.

Identification of immature spiders is extremely difficult, especially in the tropics, where even many adult specimens must be classified as morphoespecies because of insufficient taxonomic knowledge (Scharff *et al*., 2003). Even in temperate and Mediterranean areas, where the spider fauna is better known, identification of juveniles is not an easy task. However, in these geographic zones, surveys of a limited area for an extensive period provide reliable data on adult species composition and so could lead to the identification of many juveniles (e. g. Toft, 1976). Besides, ease of identification of immature stages varies with spider family and depends, mainly, on morphological distinctiveness and diversity of the taxa in the study area; e. g., juveniles of Gnaphosidae, Lycosidae or Linyphidae are much more difficult to identify than those of the two families treated in this paper. When working with the entire spider fauna, using just identifiable juveniles will favour low diversity families versus high diversity ones, which probably contain the rarest species. This may introduce taxonomic and spatial bias when comparative studies are undertaken between families or sites, respectively. Much more research on juvenile identification, perhaps involving material from ecological studies, where large numbers of specimens are collected facilitating the match between immature and adult stages, should be carried out. As pointed out by Grove (2003), to maintain biodiversity project data integrity, juveniles and adults must be stored separately, in order to analyze them as advances in their identification are achieved.

## ACKNOWLEDGMENTS

The comments of an anonymous referee greatly improved this manuscript. This paper has been supported by a Fundación BBVA Project, a MEC Project (CGL2004-04309), and also by a PhD Museo Nacional de Ciencias Naturales/C.S.I.C./Comunidad de Madrid grant.

## LITERATURE CITED

Abraham, B. J. (1983) Spatial and temporal patterns in a sagebrush steppe spider community (Arachnida, Araneae). *Journal of Arachnology*, **11**, 31-50.

Baltanás, A. (1992) On the use of some methods for the estimation of species richness. *Oikos*, **65**, 484-492.

Borges, P. A. V. & Brown, V. K. (2003) Estimating species richness of arthropods in Azorean pastures: the adequacy of suction sampling and pitfall trapping. *Graellsia*, **59**, 7-24.

Boulinier, T., Nichols, J. D., Sauer, J. R., Hines, J. E. & Pollock, K. H. (1998) Estimating species richness: the importance of heterogeneity in species detectability. *Ecology*, **79**, 1018-1028.

Brose, U., Martínez, N. D. & Williams, R. J. (2003) Estimating species richness: sensitivity to sample coverage and insensitivity to spatial patterns. *Ecology*, **84**, 2364-2377.

Cardoso, P. (2004) *The use of arachnids (Class Arachnida) in biodiversity evaluation and monitoring of natural areas.* Ph. D. thesis, Universidad de Lisboa, Portugal.

Cardoso, P., Silva, I., de Olivera, N. G. & Serrano, A. R. M. (2004) Indicador taxa of spider (*Araneae*) diversity and their efficiency in conservation. *Biological Conservation*, **120**, 517-524.

Chao, A. (1987) Estimating the population size for capture-recapture data with unequal catchability. *Biometrics*, **43**, 783-791.

Chiarucci, A., Enrigut, N. J., Perry, G. L. W., Miller, B. P. & Lamont, B. B. (2003) Performance of nonparametric species richness estimators in a high diversity plant community. *Diversity and Distributions*, **9**, 283-295.

Churchill, T. B. & Arthur, J. M. (1999) Measuring spider richness: effects of different sampling methods and spatial and temporal scales. *Journal of Insect Conservation*, **3**, 287-295.

Coddington, J. A., Young, L. H. & Coyle, F. A. (1996) Estimating spider species richness in a southern Appalachian cove hardwood forest. *Journal of Arachnology*, **24**, 111-128.

Colwell, R. K. (1997) *EstimateS: Statistical Estimation of Species Richness and Shared Species from Samples (Software and User´s Guide), Version 5.0.1*, available in http://viceroy.eeb.uconn.edu/estimates

Colwell, R. K. & Coddington, J. A. (1994) Estimating terrestrial biodiversity through extrapolation. *Philosophical Transactions of the Royal Society (series B)*, **345**, 101-118.

Dobyns, J. R. (1997) Effects of sampling intensity on the collection of spider (Araneae) species and the estimation of species richness. *Environmental Entomology*, **26**, 150-162.

Fagan, W. F. & Kareiva, P. M. (1997) Using compiled species lists to make biodiversity comparisons among regions: a test case using Oregon butterflies. *Biological Conservation*, **80**, 249-259.

Flather, C. H. (1996) Fitting species-accumulation functions and assessing regional land use impacts on avian diversity. *Journal of Biogeography*, **23**,155-168.

French, K. (1999) Spatial variability in species composition in birds and insects. *Journal of Insect Conservation*, **3**, 183-189.

Gotelli, N. J. & Collwell, R. K. (2001) Quantifying biodiversity: procedures and pitfalls in the measurement and comparison of species richness. *Ecology Letters*, **4**, 379-391.

Grove, S. J. (2003) Maintaining data integrity in insect biodiversity assessment projects. *Journal of Insect Conservation*, **7**, 33-44.

Heltshe, J. F. & Forrester, N. E. (1983) Estimating species richness using the Jacknife procedure. *Biometrics*, **39**, 1-11.

Jerardino, M., Urones, C. & Fernández, J. L. (1991) Datos ecológicos de las arañas epigeas en dos bosques de la región mediterránea. *Orsis*, **6**, 141-157.

Jiménez-Valverde, A. & Lobo, J. M. (2005) Determining a combined sampling procedure for a reliable estimation of Araneidae and Thomisidae assemblages (Araneae). *Journal of Arachnology*, **33**, 33-42.

Jiménez-Valverde, A., Jiménez Mendoza, S., Martín Cano, J & Munguira, M. L. (2006) Comparing relative model fit of species accumulation functions to local Papilionoidea and Hesperioidea butterfly inventories of Mediterranean habitats. *Biodiversity and Conservation*, **15**, 177-190.

Keating, K. A. & Quinn, J. F. (1998) Estimating species richness: the Michaelis-Menten model revised. *Oikos*, **81**, 411-416.

Koch, S. O., Chown, S. L., Davis, A. L. V., Endrödy-Younga, S. & van Jaarsveld, A. S. (2000) Conservation strategies for poorly surveyed taxa: a dung beetle (Coleoptera, Scarabaeidae) case study from southern Africa. *Journal of Insect Conservation*, **4**, 45-56.

Kotze, D. J. & Samways, M. J. (1999) Support for the multi-taxa approach in biodiversity assessment, as shown by epigaeic invertebrates in an Afromontane forest archipielago. *Journal of Insect Conservation*, **3**, 125-143.

Kremen, C., Colwell, R. K., Erwin, T. L., Murphy, D. D., Noss, R. F. & Sanjayan, M. A. (1993) Terrestrial arthropod assemblages: their use in conservation planning. *Conservation Biology*, **7**, 796-808.

Kuntner, M. & Baxter, I. H. (1997) A preliminary investigation of spider species richness in an eastern Slovenian broadleaf forest. *Proceedings of the 16th European Colloquium of Arachnology*, 173-182.

Landau, D., Prowell, D. & Carlton, C. E. (1999) Intensive versus long-term sampling to assess lepidopteran diversity in a southern mixed mesophytic forest. *Annals of the Entomological Society of America*, **92**, 435-441.

Legendre, P. & Legendre, L. (1998) *Numerical Ecology.* Elsevier, Amsterdam.

León-Cortés, J. L., Soberón-Mainero, J. & Llorente-Bousquets, J. (1998) Assessing completeness of Mexican sphinx moth inventories through species accumulation functions. *Diversity and Distributions*, **4**, 37-44.

Lobo, J. M. & Favila, E. (1999) Different ways of constructing octaves and their consequences on the prevalence of the bimodal species abundance distribution. *Oikos*, **87**, 321-326.

Maelfait, J. –P. & Desender, K. (1990) Possibilities of short-term Carabid sampling for site assessment studies. In *The Role of Ground Beetles in Ecological and Environmental Studies*, ed. N. E. Stork, pp. 217-225. Intercept, Andover, Hampshire.

Magurran, A. E. (1988) *Ecological Diversity and its Measurement.* Princeton University Press, New Jersey.

Melo, A. S., Pereira, R. A. S., Santos, A. J., Shepherd, G. J., Machado, G., Medeiros, H. F. & Azuaya, R. J. (2003) Comparing species richness among assemblages using sample units: why not use extrapolation methods to standardize different sample sizes? *Oikos*, **101**, 398-410.

Molina, J. M. (1989) Dinámica temporal de los ropalóceros de la sierra del norte de Sevilla ( *Lepidoptera: Papilionoide* et *Hesperioidea*). *Ecología*, **3**, 323-329.

Moreno, C. E. & Halffter, G. (2000) Assessing the completeness of bat biodiversity inventories using species accumulation curves. *Journal of Applied Ecology*, **37**, 149-158.

Norris, K. C. (1999) Quantifying change through time in spider assemblages: sampling methods, indices and sources of error. *Journal of Insect Conservation*, **3**, 309-325.

Palmer, M. W. (1990) The estimation of species richness by extrapolation. *Ecology*, **71**, 1195-1198.

Palmer, M. W. (1991) Estimating species richness: the second-order jacknife reconsidered. *Ecology*, **72**, 1512-1513.

Petersen, F. T. & Meier, R. (2003) Testing species-richness estimation methods on single-sample collection data using the Danish Diptera. *Biodiversity and Conservation*, **12**, 667-686.

Petersen, F. T., Meier, R. & Larsen, M. N. (2003) Testing species richness estimation methods using museum label data on the Danish Asilidae. *Biodiversity and Conservation*, **12**, 687-701.

Peterson, A. T. & Slade, N. A. (1998) Extrapolating inventory results into biodiversity estimates and the importance of stopping rules. *Diversity and Distributions*, **4**, 95-105.

Preston, F. W. (1948) The commoness, and rarity, of species. *Ecology*, **29**, 254-283.

Preston, F. W. (1962) The canonical distribution of commoness and rarity. *Ecology*, **43**, 185-215.

Reid, W. V. (1998) Biodiversity hotspots. *Trends in Ecology and Evolution*, **13**, 275-279.

Riecken, U. (1999) Effects of short-term sampling on ecological characterization and evaluation of epigeic spider communities and their habitats for site assessment studies. *Journal of Arachnology*, **27**, 189-195.

Rohlf, F. J. (2000) *NTSYSpc numerical taxonomy and multivariate analysis system version 2.1*. Exeter Software, Setauket, NY.

Sætersdal, M. Gjerde, I., Blom, H. H., Ihlen, P. G., Myrseth, E. W., Pommeresche, R., Skartveit, J., Solhøy, T. & Aas, O. (2003) Vascular plants as a surrogate species group in complementary site selection for bryophytes, macrolichens, spiders, carabids, staphylinids, sanils, and wood living polypore fungi in a northern forest. *Biological Conservation*, **115**, 21-31.

Scharff, N., Coddington, J. A., Griswold, C. E., Hormiga, G. & Bjørn, P. de P. (2003) When to quit? Estimating spider species richness in a northern european deciduous forest. *Journal of Arachnology*, **31**, 246-273.

Shapiro, A. M. (1975) The temporal component of butterfly species diversity. In *Ecology and evolution of communities,* eds. M. L. Cody & J. M. Diamond, pp. 181 195. Harvard University Press, Cambridge, MA.

Smith, E. P. & van Belle, G. (1984) Nonparametric estimation of species richness. *Biometrics*, **40**, 119-129.

Soberón, J. & Llorente, B. J. (1993) The use of species accumulation functions for the prediction of species richness. *Conservation Biology*, **7**, 480-488.

Sørensen, L. L., Coddington, J. A. & Scharff, N. (2002) Inventorying and estimating subcanopy spider diversity using semiquantitative sampling methods in an Afromontane forest. *Environmental Entomology*, **31**, 319-330.

Sørensen, L. L. (2004) Composition and diversity of the spider fauna in the canopy of a montane forest in Tanzania. *Biodiversity and Conservation*, **13**, 437-452.

StatSoft (2001) *STATISTICA (data analysis software system and user´s manual). Version 6.* StatSoft, Inc., Tulsa, OK.

Toft, S. (1976) Life-histories of spiders in a Danish beech wood. *Natura Jutlandica*, **19**, 5-40.

Tokeshi, M. (1993) Species abundance patterns and community structure. *Advances in Ecological Research*, **24**, 111-186.

Toti, D. S., Coyle, F. A. & Miller, J. A. (2000) A structured inventory of Appalachian grass bald and heath bald spider assemblages and a test of species richness estimator performance. *Journal of Arachnology*, **28**, 329-345.

Urones, C. & Puerto, A. (1988) Ecological study of the Clubionoidea and Thomisoidea (Araneae) in the Spanish Central System. *Revue Arachnologique*, **8**, 1-32.

Walther, B. A. & Martin, J.-L. (2001) Species richness estimation of bird communities: how to control for sampling effort? *Ibis*, **143**, 413-419.

Willott, S. J. (2001) Species accumulation curves and the measure of sampling effort. *Journal of Applied Ecology*, **38**, 484-486.

Wise, D. H. (1993) *Spiders in ecological webs.* Cambridge University Press, New York.

# UN MÉTODO SENCILLO PARA SELECCIONAR PUNTOS DE MUESTREO CON EL OBJETO DE INVENTARIAR TAXONES HIPERDIVERSOS: EL CASO PRÁCTICO DE LAS FAMILIAS ARANEIDAE Y THOMISIDAE (ARANEAE) EN LA COMUNIDAD DE MADRID, ESPAÑA

**RESUMEN.** Elaborar estrategias de conservación eficaces requiere poseer información faunística bien repartida a lo largo del espectro de condiciones ambientales de una región. Ello es posible si la toma de datos se ha realizado mediante el desarrollo de protocolos de muestreo bien diseñados, eficientes y específicos para cada grupo biológico. En este trabajo se presenta una metodología sencilla para la selección de puntos de muestreo, especialmente útil en el caso de grupos hiperdiversos en los que la recolección de información faunística requiere un esfuerzo notable y la determinación taxonómica del material colectado no es posible en el campo. El método se basa en la clasificación de las unidades territoriales de una región de acuerdo a los valores de una serie de variables ambientales y espaciales, previamente seleccionadas por su conocida influencia sobre la distribución del grupo de organismos considerado y compiladas en un Sistema de Información Geográfica. Tras definir la superficie y el número de unidades territoriales en las que es posible obtener inventarios fiables, se utiliza la estrategia de agrupamiento *k-means* a fin de obtener una clasificación de la región en tantas subregiones como unidades territoriales se vayan a muestrear. Dentro de cada subregión, la unidad territorial puede ser seleccionada teniendo en cuenta diversos criterios, como su distancia espacio-ambiental al centroide de la subregión, su facilidad de acceso o el volumen de información previamente existente. Se ofrece un ejemplo

práctico de esta metodología con las familias de arañas Araneidae y Thomisidae en la Comunidad de Madrid.

**Palabras clave:** inventarios biológicos, protocolos de muestreo, clasificación espacio-ambiental, algoritmo *k-means*, Sistemas de Información Geográfica, Araneidae, Thomisidae, Comunidad de Madrid

**Este capítulo ha sido publicado en:**

JIMÉNEZ-VALVERDE, A. & LOBO, J. M. (2004). Un método sencillo para seleccionar puntos de muestreo con el objeto de inventariar taxones hiperdiversos: el caso práctico de las familias *Araneidae* y *Thomisidae (Araneae)* en la Comunidad de Madrid (España). *Ecología,* **18**, 297-308.

## INTRODUCCIÓN

Actualmente, el incesante aumento de población y el ritmo insostenible de consumo de recursos naturales está provocando la pérdida de diversidad biológica de forma acelerada, siendo éste uno de los problemas ambientales más graves (Wilson, 1999; Myers, 2003). Para poder abordarlo y elaborar estrategias de conservación eficientes es necesario disponer de información corológica precisa y no sesgada (Williams *et al.*, 2002). Sin embargo, la frecuente ausencia de datos dificulta esta labor, especialmente cuando tratamos con grupos hiperdiversos como son los artrópodos. La pérdida de apoyo económico y político que la taxonomía sufre y ha sufrido en favor de otras disciplinas más competitivas (Charles & Godfray, 2002), la consecuente ausencia de especialistas en muchos grupos (Martín-Piera & Lobo, 2000; Valdecasas & Camacho, 2003), los sesgos en la distribución geográfica del conocimiento taxonómico debidos a la preponderancia de colectas dirigidas hacia los lugares de residencia de los especialistas, las áreas visualmente atractivas, o los enclaves reconocidos por su riqueza en especies (García-Barros & Munguira, 1999; Martín & Gurrea, 1999; Dennis & Thomas, 2000; Reddy & Dávalos, 2003) son, entre otros, factores que imposibilitan conocer cuántas especies se encuentran en una localidad determinada y, por supuesto, cuál es la identidad de esas especies y cuál la distribución geográfica de cada una de ellas. Es indudable que disponer de estos conocimientos podría cambiar las estrategias de conservación, actualmente centradas en vertebrados y plantas.

Hoy en día, distintas técnicas estadísticas y los Sistemas de Información Geográfica permiten realizar modelos predictivos de distribución, tanto de especies concretas como de los diferentes atributos que representan la biodiversidad (riqueza específica, rareza, endemicidad, etc.), plasmando los resultados en un mapa extrapolado

(Guisan & Zimmermann, 2000; Lobo, 2000; Hirzel *et al.*, 2002; Hortal & Lobo, 2002; Zaniewski *et al.*, 2002; Peterson & Kluza, 2003; Store & Jokimäki, 2003; Wang *et al.*, 2003, entre otros). Estas técnicas de modelización juegan, cada vez más, un papel esencial a la hora de desarrollar estrategias de conservación (Andriamampianina *et al.*, 2000; Peterson *et al.*, 2000; Bailey *et al.*, 2002; Schadt *et al.*, 2002; Suárez-Seoane *et al.*, 2002; Barbosa *et al.*, 2003) y son, tal vez, la única manera fiable de identificar a corto plazo las áreas de mayor diversidad en territorios insuficientemente muestreados, al objeto de considerar esa información en los planes de gestión del territorio.

Sin embargo, para poder construir estos mapas predictivos es necesario disponer de unos pocos inventarios fiables que recojan el máximo rango posible de variación del taxón o atributo en cuestión en el territorio seleccionado, por lo que un buen diseño de muestreo resulta esencial. Existen varias estrategias para seleccionar los puntos de muestreo que pueden o no considerar la información ambiental del territorio. Las aproximaciones más sencillas, como el muestreo aleatorio y el muestreo sistemático (ver, por ejemplo, Southwood & Henderson, 2000) buscan ubicar las localidades de muestreo independientemente de las condiciones ambientales. Por el contrario, el muestreo de tipo estratificado trata de asegurar que las colectas se efectúen en la mayor variedad de ambientes posible, subdividiendo el territorio en regiones ambientalmente homogéneas (Austin & Heyligers, 1989, 1991; Guisan & Zimmermann, 2000). Un caso particular de muestreo estratificado es el método GRADSECT, el cual pretende encontrar el gradiente de localidades que maximiza la variabilidad ambiental del territorio y ha demostrado ser más eficiente que los métodos aleatorios o sistemáticos cuando se trata de conseguir una muestra representativa de localidades de un territorio (Wessels *et al.*, 1998). Otras aproximaciones más complejas y eficaces que tienen en cuenta la información ambiental se basan en el concepto de diversidad ambiental (ED,

"environmental diversity") y en métodos de selección (como el de la *p-media*) capaces de determinar el emplazamiento óptimo de una localidad en un territorio o espacio ambiental multidimensional (Faith & Walker, 1996; Araújo *et al.*, 2001; Ferrier, 2002).

Aunque los últimos métodos mencionados sean eficaces y adecuados para identificar la ubicación de los puntos de colecta, requieren la utilización de técnicas estadísticas relativamente complejas y poco accesibles y, a nuestro juicio, adolecen de dos inconveniente principales. Por una parte, no se interesan ni ofrecen ninguna indicación sobre las variables ambientales que deben considerarse a la hora de regionalizar el territorio (Mohler, 1983; Hirzel & Guisan, 2002) y, sobre todo, no consideran aquellas variables que más influyen sobre los organismos a colectar. Es decir, ofrecen un panorama ambiental condicionado por nuestra visión antropocéntrica y poco relacionado con el punto de vista de los organismos. Por otra, no tienen en cuenta la estructura espacial del territorio y ésta, cuando las variables ambientales consideradas no son capaces de hacerlo, pueden explicar la variación espacial en la diversidad biológica frecuentemente debida a factores geográficos o históricos únicos e irrepetibles (Legendre & Legendre, 1998; Lobo, 2000; Araújo *et al.,* 2001, 2003).

El inventario de taxones hiperdiversos, grupos sobre los que normalmente se dispone de escasa información, suele ser arduo y costoso, tanto en términos económicos como humanos y de tiempo. La metodología que proponemos en este trabajo se basa en el análisis de agrupamiento y está diseñada para regionalizar un territorio de acuerdo al esfuerzo máximo que es posible realizar en el trabajo de campo y teniendo en cuenta, tanto la variabilidad ambiental que afecta al grupo taxonómico elegido, como la posición espacial de las localidades. El esfuerzo de colecta está, por tanto, definido previamente y constituye el criterio inicial y realista con el que se define la búsqueda del número de unidades territoriales a muestrear. El método, por tanto, maximiza la

variabilidad espacio-ambiental recogida en función del esfuerzo. Como ejemplo para ilustrar el procedimiento propuesto se emplea diversa información ambiental y espacial de la Comunidad de Madrid, al objeto de delimitar las localidades necesarias para efectuar un muestreo de dos familias de arañas (Araneidae y Thomisidae) en este territorio.

## ESCALA DE TRABAJO Y ESFUERZO DE MUESTREO

El primer paso del proceso consiste en decidir la extensión del territorio y la resolución o tamaño de celda a la que vamos a trabajar, ya que los patrones de diversidad y los factores que determinan la riqueza específica están influenciados por estas variables (Wiens *et al.,* 1986; ver ejemplos en Böhning-Gaese, 1997 y Martínez *et al.*, 2003 entre otros). Evidentemente, nuestra capacidad de trabajo condicionará tanto el área total del territorio, como el tamaño y el número de las unidades territoriales en las que podemos obtener inventarios fiables (ver Blackburn & Gaston, 2002 para una discusión sobre la escala idónea de trabajo). Para determinar la resolución de trabajo resulta útil examinar la relación entre el incremento en el esfuerzo realizado y la acumulación de especies encontradas (Colwell & Coddington, 1994; Gotelli & Colwell, 2001). Las unidades de esfuerzo de muestreo pueden ser horas de observación, número de trampas, cuadrados de muestreo, etc. Otras veces son unidades más complejas como las empleadas en Coddington *et al.* (1996), Toti *et al.* (2000) y Jiménez-Valverde & Lobo (2005), donde cada unidad de muestreo está constituida por un conjunto complementario de métodos de colecta diferentes, cada uno utilizado durante un tiempo determinado y, en ocasiones, por personas diferentes Al principio, la adición de especies al inventario se produce rápidamente y, por tanto, la pendiente de la curva comienza

siendo elevada pero, a medida que el muestreo avanza, sólo se adicionan las especies raras, descendiendo paulatinamente la pendiente de la curva de acumulación. El momento en el que esta pendiente desciende a cero (es decir, cuando se alcanza la asíntota) corresponde, teóricamente, al número total de especies que podemos encontrar en la zona estudiada con los métodos utilizados y durante el periodo en el que se llevó a cabo el muestreo. Aleatorizando el orden de entrada de las unidades muestrales en numerosas ocasiones, es posible obtener el número de especies promedio para cada cantidad de esfuerzo (Colwell, 2000). Estos valores pueden entonces ajustarse a distintos tipos de funciones (Soberón & Llorente, 1993; Colwell & Coddington, 1994), permitiendo calcular tanto el número total de especies que supuestamente es posible colectar, como la tasa de incremento en el número de especies a un esfuerzo concreto valorando, así, el grado de precisión del inventario en cada momento (ver Jiménez-Valverde & Hortal, 2003). Hay que tener en cuenta que el esfuerzo de muestreo necesario para lograr inventarios completos variará en función de la complejidad estructural del hábitat de tal manera que, en general, los lugares más complejos necesitarán una inversión de esfuerzo mayor. Es necesario, por tanto, construir diferentes curvas para lugares de distinta complejidad estructural y alejarnos de la idea preconcebida según la cual es necesario realizar un esfuerzo de colecta idéntico en cada una de las localidades seleccionadas.

Comparando curvas de acumulación realizadas con los datos provenientes de unidades territoriales de distinto tamaño, será posible estimar la resolución adecuada a la que podemos obtener inventarios fiables, determinado así el número de unidades territoriales o localidades a muestrear y, consecuentemente, la extensión idónea de nuestro territorio de colecta. Para ello, debe evaluarse la capacidad de trabajo que podemos desplegar en la realización del inventario, así como otras limitaciones que

puedan surgir (económicas, disponibilidad de tiempo, etc.). La delimitación del número de puntos de colecta que vayamos a ser capaces de muestrear constituye la fase fundamental de este proceso metodológico, ya que determina el número de subregiones en las que, mediante el análisis de agrupamiento, se dividirá el territorio elegido.

## CLASIFICACIÓN ESPACIO-AMBIENTAL Y SELECCIÓN DE LOS PUNTOS DE MUESTREO

El análisis de agrupamiento es un método estadístico multivariante de clasificación de datos. Engloba una serie de técnicas algorítmicas destinadas a clasificar observaciones en grupos homogéneos en función de una serie de variables. Entre las diversas técnicas de agrupamiento, el método *k-means* se basa en un algoritmo heurístico de clasificación no jerárquica que, partiendo de un número concreto de centroides o grupos previamente definidos por el usuario, trata de seleccionar una configuración que minimice la dispersión de los valores de las variables utilizadas en la clasificación dentro de cada grupo, maximizando la variación entre los grupos (Legendre & Legendre, 1998). Nosotros hemos utilizado este procedimiento de agrupación para delimitar un número de subregiones espacio-ambientales, idéntico al número de puntos de colecta que vayamos a ser capaces de muestrear.

Las variables ambientales que pueden emplearse para la clasificación de las unidades territoriales son elegidas en función de su influencia sobre la distribución del grupo taxonómico de estudio. Cuando sea posible, ello puede determinarse utilizando datos biológicos y ambientales a la escala y extensión espacial elegida y, por ejemplo, realizando un análisis de regresión múltiple con esa información. Sin embargo, cuando no pueda disponerse de esa información, la selección de las variables ambientales puede

basarse en el conocimiento general sobre los factores ambientales que condicionan la distribución del grupo elegido. Si delimitamos conjuntos de localidades o subregiones utilizando únicamente variables ambientales, es probable que se agrupen por su similitud unidades territoriales que, aunque estén alejadas en el espacio, presenten condiciones ambientales similares. Sin embargo, esas zonas disjuntas de la misma subregión ambiental, pueden albergar elementos faunísticos o florísticos diferentes debido a la actuación de factores contingentes únicos e irrepetibles (Lobo, 1997). Al objeto de promover la aparición de subregiones homogéneas ambientalmente y, a la vez, espacialmente continuas, resulta conveniente incluir la latitud y la longitud dentro del conjunto de las variables consideradas, o incluso, los nueves términos de una función polinomial de tercer grado de estas dos variables (*Trend Surface Analysis,* ver Legendre & Legendre, 1998). Toda esta información espacio-ambiental puede compilarse mediante un Sistema de Información Geográfica (Johnston, 1998). Antes de realizar el análisis de agrupamiento es necesario estandarizar las variables seleccionadas para evitar sesgos debidos a la diferencia en las unidades de medida de las distintas variables. Si el número de variables ambientales es elevado es probable que existan correlaciones entre ellas; en ese caso puede resultar conveniente reducir su número mediante alguna técnica de ordenación que posibilite la creación de nuevas variables ortogonales entre si (por ejemplo, Análisis de Componentes Principales o Análisis de Coordenadas Principales; ver Legendre & Legendre, 1998). Sin embargo, siempre se ha de tener en cuenta que estos métodos provocan una pérdida en la variabilidad espacio-ambiental y que la ordenación en el espacio reducido que representan estas nuevas variables puede no representar fielmente la información de partida.

Una vez realizado el análisis de agrupamiento, es necesario elegir las unidades territoriales que mejor representan cada una de las subregiones. En estos conjuntos de

localidades o subregiones, el área que ocupa cada agrupación esta formada por localidades dispuestas a lo largo de un gradiente espacio-ambiental, en el cual algunas se encuentran más cerca de los valores espacio-ambientales promedios de cada subregión que otras. De esta manera, para que quede representada lo más fielmente posible la variabilidad espacio-ambiental de la región, deberían seleccionarse aquellas unidades territoriales con valores más cercanos a los centroides de cada agrupación. Podemos emplear otros criterios de manera jerárquica e iterativa hasta lograr una selección definitiva de las localidades de muestreo. Estos criterios pueden ser la facilidad de acceso o la existencia de información biológica previamente existente.

## EJEMPLO PRÁCTICO: LAS FAMILIAS DE ARAÑAS ARANEIDAE Y THOMISIDAE EN LA COMUNIDAD DE MADRID

**El grupo taxonómico elegido. —** El orden Araneae es uno de los más diversificados. Hasta el momento hay descritas alrededor de 36000 especies de arañas en todo el mundo, aunque se estima que deben existir entre 60000 y 170000 (Coddington & Levi, 1991). Son depredadores generalistas abundantes y ubiquistas y, por ello, tienen gran importancia en los sistemas ecológicos (Wise, 1993). Sin embargo, todavía es necesario dedicar un gran esfuerzo hasta lograr desarrollar protocolos estandarizados de muestreo de la fauna aracnológica, siendo además imprescindible realizar estudios sobre su ecología y distribución, a fin de propiciar que estas especies adquieran la relevancia que les corresponde en los planes de gestión y conservación (New, 1999). La familia Araneidae es una de las más exitosas (aproximadamente 2600 especies descritas; Foelix, 1996). Son arañas que construyen telas orbiculares para la captura de sus presas y, por tanto, la estructura de la vegetación parece ser el parámetro

más importante a la hora de determinar su presencia (Wise, 1993). Las arañas de la familia Thomisidae no emplean tela para la captura de sus presas, sino que permanecen al acecho sobre hojas y flores, pasando inadvertidas gracias a su coloración críptica. Algunos géneros, como *Xysticus* y *Ozyptila*, son eminentemente edáficos, capturando a sus presas entre la hojarasca y la vegetación herbácea.

El grado de conocimiento sobre la distribución de las arañas de la Península Ibérica es sumamente limitado debido a una pobre tradición aracnológica. Únicamente la Comunidad de Aragón cuenta con un catálogo reciente de su fauna aracnológica (Melic, 2000), mientras que el resto de los catálogos ibéricos son, en realidad, compendios de citas antiguas, muchas de ellas dudosas o erróneas (Melic, 2001). Además, existen pocos estudios que hayan sido realizados durante un periodo prolongado de tiempo y que hayan empleado distintas técnicas complementarias de captura, por lo que, generalmente, la información disponible presenta importantes sesgos. Como consecuencia, son escasas las localidades de la Península que cuentan con un inventario más o menos completo de su fauna aracnológica. Urge, pues, realizar estudios faunísticos en la Península Ibérica pero, a fin de rentabilizar el esfuerzo a realizar, estos deberían hacerse siguiendo protocolos bien diseñados y estandarizados.

**La región de estudio. —** La Comunidad de Madrid, el territorio al que se ha aplicado la metodología propuesta, se sitúa en el centro de la Península Ibérica y presenta una superficie de más de 8000 $km^2$. El relieve distingue tres grandes zonas situadas en un gradiente NO-SE: la Sierra, la Zona de Transición y las Llanuras del Tajo. La Sierra forma parte del Sistema Central y presenta una orientación NE-SO. Está formada casi en su totalidad por rocas plutónicas y metamórficas, con pequeños afloramientos calizos en su tercio norte. En las otras dos unidades de relieve afloran materiales producto de la erosión y posterior sedimentación, predominando calizas,

margas, yesos, arenas y arcillas. En el sector Guadarrámico de la Sierra se encuentra la mayor altura de la Comunidad, el Pico de Peñalara con 2430 m. y en el valle del río Alberche la menor (434 m). La altitud media de este territorio es de unos 800 m. El clima es de tipo continental con influencia mediterránea, con precipitaciones anuales que oscilan entre los 350 mm y los 2000 mm anuales. La Comunidad de Madrid es un territorio heterogéneo que se sitúa fitosociológicamente dentro de la región Mediterránea, estando en ella representados los pisos Mesomediterráneo, Supramediterráneo, Oromediterráneo y Crioromediterráneo.

**Escala de trabajo y esfuerzo de muestreo. —** Jiménez-Valverde & Lobo (2005) estudian la efectividad de distintos métodos para la captura de araneidos y tomísidos y concluyen que una combinación de, al menos, tres métodos complementarios permite prospectar cuadrículas de 1 $km^2$ obteniendo inventarios fiables. Para realizar un muestreo adecuado de las especies de estas familias en una cuadrícula de 1 $km^2$ relativamente rica en especies de la Comunidad de Madrid, los autores llevaron a cabo 20 unidades de esfuerzo de muestreo en parcelas de 400 $m^2$ dispuestas aleatoriamente sobre el terreno. Cada unidad de muestreo estaba definida como 4 trampas de caída actuando durante 48 horas, mangueo de la vegetación herbácea y subarbustiva, y batido de la cubierta arbustiva y arbórea durante 15 minutos respectivamente (para una descripción detallada ver Jiménez-Valverde & Lobo, 2005). Las curvas de acumulación de especies obtenidas por los autores se aproximaban bastante a la asíntota comprobando que, tras realizar el ajuste de la función de Clench, recogían alrededor del 80% del total de especies estimado. El ingente trabajo de muestreo a realizar impedía obtener, con el esfuerzo de trabajo disponible, inventarios fiables de estas especies a resoluciones mayores. Debido a ello, se estimó que podían muestrearse 15 unidades territoriales de 1 $km^2$ en el tiempo y con los recursos

disponibles. Una recopilación de toda la información corológica disponible en las colecciones y la bibliografía sobre los araneidos y tomísidos de la Comunidad de Madrid, permitió comprobar que una gran parte de las citas disponibles no eran aptas para este estudio, debido a que poseían una precisión espacial mucho mayor (generalmente, cuadrículas de 10 $km^2$).

**Selección, obtención y preparación de la información ambiental**. **—** A falta de información taxonómica y corológica fiable de distintas localidades, la relevancia diferencial de las distintas variables ambientales se estimó teniendo en cuenta la información general conocida sobre estos grupos. De acuerdo a las referencias bibliográficas, cuatro son los factores generalmente admitidos como determinantes de la distribución de las arañas: la estructura de la vegetación (Hatley & Macmahon, 1980; Robinson, 1981; Rypstra, 1986; Urones & Puerto, 1988; Döbel *et al*., 1990; Uetz, 1991; Wise, 1993; Downie *et al.,* 1995; Balfour & Rypstra, 1998; Downie *et al.*, 2000; Borges & Brown, 2001; Urones & Majadas, 2002), la humedad (Coulson & Butterfield, 1986; Rushton *et al.*, 1987; Rushton & Eyre, 1992; Bonte *et al.,* 2002), la temperatura (Rypstra, 1986) y la altitud (Rushton & Eyre, 1992; Urones & Puerto, 1988). Además de estas variables, se tuvo también en cuenta la litología (sustrato calizo o silíceo), ya que ésta va a determinar la permeabilidad del sustrato y, a su vez, la distinta disponibilidad de agua.

Mediante un Sistema de Información Geográfica (Idrisi 32, Clark Labs, 2000a) se crearon cinco capas temáticas, una para cada variable anteriormente mencionada. La estructura de la vegetación se obtuvo reclasificando los tipos de uso del suelo del programa CORINE (European Environment Agency, 1996) en tres categorías que representaban niveles de complejidad estructural creciente: pastos, vegetación arbustiva y formaciones arbóreas. Como la resolución espacial de esta información cartográfica

digital se encontraba disponible en píxeles de 250 m de lado, hubo que adecuar ésta al tamaño seleccionado de las unidades territoriales (cuadrículas UTM de 1 km de lado). La información climática se obtuvo utilizando los valores promedio (30 años) de temperatura media mensual y precipitación total anual provenientes de 41 estaciones climatológicas del centro peninsular (Ministerio de Agricultura, Pesca Y Alimentación, 1986), e interpolando mediante medias móviles una cartografía de estas variables a la resolución requerida. La altitud media de cada cuadrícula de 1 km se obtuvo a partir de un modelo digital de elevación del planeta (Clark Labs, 2000b). Para la litología se definieron dos categorías: suelos básicos (calizas, yesos y rocas afines) y suelos ácidos (rocas silíceas y sedimentos derivados) a partir de un mapa litológico en soporte papel (ITGE, 1988) que, una vez digitalizado, fue reclasificado a una resolución de cuadrículas UTM de 1 $km^2$. En resumen, para realizar la regionalización espacio-ambiental de la Comunidad de Madrid se emplearon cinco variables ambientales, una variable cualitativa multinomial con tres estados (estructura de la vegetación, VEG), una variable cualitativa binomial con dos estados (litología, LIT), tres variables cuantitativas continuas (altitud, ALT; temperatura media anual, TEMP y precipitación media anual, PRECP). A éstas se sumaron dos variables espaciales: la latitud (LAT) y la longitud (LONG) central de dada cuadrícula UTM de 1 $km^2$. Por último, mencionar que no fueron consideradas aquellas cuadrículas en las que dominaban los usos del suelo urbano, los cultivos de cualquier tipo, las láminas de agua, los arenales y la roca desnuda.

## CLASIFICACIÓN ESPACIO-AMBIENTAL DE LA COMUNIDAD DE MADRID Y SELECCIÓN DE LOS PUNTOS DE MUESTREO

Tras estandarizar las variables espaciales a media cero (Legendre & Legendre, 1998) y las variables ambientales continuas a media cero y desviación estándar 1, se regionalizó la Comunidad de Madrid en 15 subregiones mediante un análisis de agrupamiento de la *k-media* (Fig. 1) eligiendo, como procedimiento para seleccionar los centroides iniciales, la maximización de la distancia entre ellos (StatSoft, 2001). En la Tabla 1 se muestran las principales características ambientales y geográficas de cada subregión. Las unidades territoriales en las que efectuar el muestreo dentro de cada subregión fueron seleccionadas atendiendo, primero, a la menor distancia al centroide del cluster correspondiente y, posteriormente, según su cercanía a la red de carreteras (Fig. 1).

Tabla 1.- Características ambientales y geográficas de cada una de las 15 subregiones de la Comunidad de Madrid, delimitadas mediante un análisis de agrupamiento *k-means* y la información de las variables que se detallan a continuación. ALT, altitud media en metros; PRECP, precipitación total anual en mm; TEMP, temperatura media anual en ºC; VEG, estructura de vegetación (1. pastos, 2. vegetación arbustiva, 3. vegetación arbórea); LIT, litología (1. suelos ácidos, 2. suelos básicos). En el caso de VEG y LIT se muestra en negrita el valor mayoritario en el caso de que la subregión posea cuadrículas con diferentes estados de la misma variable.

| Subregión | Área ($km^2$) | ALT | PRECP | TEMP | VEG | LIT |
|---|---|---|---|---|---|---|
| 1 | 298 | 1483 | 891 | 10.1 | 1, 2, **3** | **1**, 2 |
| 2 | 251 | 771 | 581 | 13.0 | **1**, 2 | **1**, 2 |
| 3 | 444 | 875 | 685 | 13.2 | 2, **3** | **1**, 2 |
| 4 | 96 | 1762 | 1309 | 7.0 | 1, **2**, 3 | 1 |
| 5 | 305 | 666 | 490 | 13.9 | 1, 2, **3** | **1**, 2 |
| 6 | 287 | 764 | 618 | 13.5 | 1, 2, **3** | 1 |
| 7 | 284 | 1525 | 715 | 11.7 | 1, 2, **3** | **1**, 2 |
| 8 | 181 | 555 | 437 | 14 | 1, **2**, 3 | 1, **2** |
| 9 | 77 | 1462 | 1116 | 8.7 | 1, **2**, 3 | 1 |
| 10 | 218 | 696 | 449 | 13.9 | 1, **3** | 1, **2** |
| 11 | 344 | 1075 | 677 | 12.2 | **2**, 3 | **1**, 2 |
| 12 | 272 | 682 | 450 | 13.9 | 2 | 1, **2** |
| 13 | 314 | 599 | 470 | 13 | 1, 2, **3** | 1 |
| 14 | 541 | 966 | 710 | 12.4 | 1 | **1**, 2 |
| 15 | 322 | 1112 | 792 | 9.8 | 1, 2, **3** | 1 |

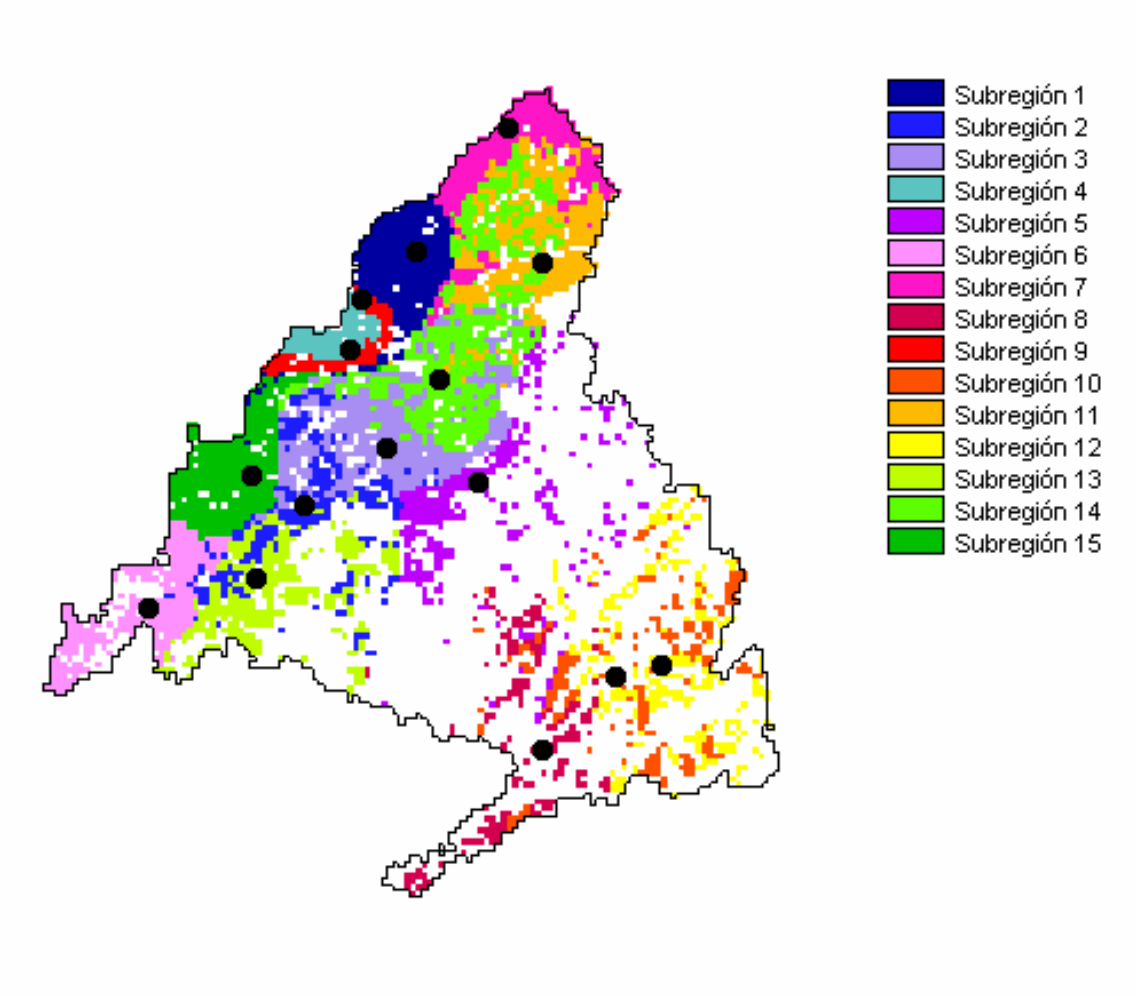



Figura 1.- División de la Comunidad de Madrid en 15 subregiones mediante un análisis de agrupamiento *k-means* y diversas variables ambientales y espaciales. Cada subregión representa un territorio con similares condiciones ambientales y constituido por unidades territoriales espacialmente contiguas. Las zonas en blanco corresponden a localidades excluidas del trabajo de inventariado debido a su uso como suelo urbano o de cultivo (ver texto). Los puntos negros son las unidades territoriales de muestreo de 1 $km^2$. Cada punto representa un cluster o subregión y es el resultado de dos criterios jerárquicos: menor distancia al centroide del cluster que representa con el objeto de que esa representación sea lo más fiel posible (ver texto), y facilidad de acceso.

## AGRADECIMIENTOS

A Joaquín Hortal por su ayuda con la información ambiental. Este trabajo se engloba dentro de los proyectos REN-2001-1136/GLO de la DGCYT ("Faunística predictiva: Análisis comparado de la efectividad de distintas metodologías y su aplicación para la selección de reservas") y 07M/0080/2002 de la Comunidad de Madrid

("Factores determinantes de la biodiversidad en la Comunidad de Madrid y predicción de su variación geográfica"). El primer autor ha podido realizar este trabajo gracias a una beca predoctoral Museo Nacional de Ciencias Naturales/ CSIC/ Comunidad de Madrid.

## REFERENCIAS BIBLIOGRÁFICAS

Andriamampianina, L., Kremen, C., Vane-Wright, D., Lees, D., & Razafimahatratra, V. (2000) Taxic richness patterns and conservation evaluation of Madagascan tiger beetles (Coleoptera: Cicindelidae). *Journal of Insect Conservation*, **4**, 109-128.

Araújo, M. B., Densham, P. J. & Humphries, C. J. (2003) Predicting species diversity with ED: the quest for evidence. *Ecography*, **26(3)**, 380-383.

Araújo, M. B., Humphries, C. J., Densham, P. J., Lampinen, R., Hagemeijer, W. J. M., Mitchell-Jones, A. J. & Gasc, J. P. (2001) Would environmental diversity be a good surrogate for species diversity? *Ecography*, **24(1),** 103-110.

Austin, M. P. & Heyligers, P. C. (1989) Vegetation survey design for conservation: gradsect sampling of forest in north-eastern New South Wales. *Biological Conservation*, **50**, 13-32.

Austin, M. P. & Heyligers, P. C. (1991) New approach to vegetation survey design: gradsect sampling. En *Nature Conservation: Cost Effective Survey and Data Analysis,* eds. C. R. Margules & M. P. Austin, pp. 31-36. CSIRO, Melbourne.

Bailey, S.-A., Haines-Young, R. H. & Watkins, C. (2002) Species presence in fragmented landscapes: modelling of species requirements at the national level. *Biological Conservation*, **108**, 307-316.

Balfour, R. A. & Rypstra, A. L. (1998) The influence of habitat structure on spider density in a no-till soybean agroecosystem. *Journal of Arachnology*, **26**, 221-226.

Barbosa, A. M., Real, R., Olivero, J. & Vargas, J. M. (2003) Otter (*Lutra lutra*) distribution modeling at two resolution scales suited to conservation planning in the Iberian Peninsula. *Biological Conservation*, **114**, 377-387.

Blackburn, T. M. & Gaston, K. J. (2002) Scale in macroecology. *Global Ecology and Biogeography*, **11**, 185-189.

Böhning-Gaese, K. (1997) Determinants of avian species richness at different spatial scales. *Journal of Biogeography*, **24**, 49-60.

Bonte, D., Baert, L. & Maelfait, J.-P. (2002) Spider assemblage structure and stability in a heterogeneous coastal dune system (Belgium). *Journal of Arachnology*, **30**, 331-343.

Borges, P. A. V. & Brown, V. K. (2001) Phytophagous insects and web-building spiders in relation to pasture vegetation complexity. *Ecography*, **24**, 68-82.

Charles, H. & Godfray, J. (2002) Challenges for taxonomy. *Nature*, **417**, 17-19.

Clark Labs. (2000a) *Idrisi 32.02. GIS software package.* Clark University.

Clark Labs. (2000b) *Gobal Change Data Archive Vol. 3. 1 km Global Elevation Model.* CD-Rom. Clark University.

Coddington, J. A. & Levi, H. W. (1991) Systematics and evolution of spiders. *Annual Review of Ecology and Systematics*, **22**, 565-592.

Coddington, J. A., Young, L. H. & Coyle, F. A. (1996) Estimating spider species richness in a southern Appalachian cove hardwood forest. *Journal of Arachnology*, **24**, 111-128.

Colwell, R. K. (2000) *EstimateS: Statistical Estimation of Species Richness and Shared Species from Samples (Software and User´s Guide), Version 6.0.* Disponibe en http://viceroy.eeb.uconn.edu/estimates

Colwell, R. K. & Coddington, J. A. (1994) Estimating terrestrial biodiversity through extrapolation. *Philosophical Transactions of the Royal Society, series B*, **345**, 101-118.

Coulson, J. C. & Butterfield, J. (1986) The spider communities on peat and upland grasslands in northern England. *Holartic Ecology*, **9**, 229-239.

Dennis, R. L. H. & Thomas, C. D. (2000) Bias in butterfly distribution maps: the influence of hot spots and recorder´s home range. *Journal of Insect Conservation*, **4**, 73-77.

Döbel, H. G., Denno, R. F. & Coddington, J. A. (1990) Spider (Araneae) community structure in an intertidal salt marsh: effects of vegetation structure and tidal flooding. *Environmental Entomology*, **19(5)**, 1356-1370.

Downie, I. S., Butterfield, J. E. L. & Coulson, J. C. (1995) Habitat preferences of sub-montane spiders in northern England. *Ecography,* **18,** 51-61.

Downie, I. S., Ribera, I., McCracken, D. I., Wilson, W. L., Foster, G. N., Waterhouse, A., Abernethy, V. J. & Murphy, K. J. (2000) Modelling populations of *Erigone atra* and *E. dentipalpis* (Araneae: Linyphiidae) across an agricultural gradient in Scotland. *Agriculture, Ecosystems and Environment*, **80**, 15-28.

European Environment Agency. (1996) *Natural Resources CD-Rom.* European Environment Agency.

Faith, D. P. & Walker, P. A. (1996) Environmental diversity: On the best-possible use of surrogate data for assessing the relative biodiversity of sets of areas. *Biodiversity and Conservation*, **5**, 399-415.

Ferrier, S. (2002) Mapping spatial pattern in biodiversity for regional conservation planning: where to from here? *Systematic Biology*, **51(2)**, 331-363.

Foelix, R. F. (1996) *Biology of spiders.* Oxford University Press, Nueva York.

García-Barros, E. & Munguira, M. L. (1999) Faunística de mariposas diurnas en España peninsular. Áreas poco estudiadas: una evaluación en el umbral del siglo XXI (Lepidoptera: Papilionidae & Hesperiidae). *SHILAP Revista de lepidopterología*, **27(106)**, 189-202.

Gotelli, N. J. & Colwell, R. K. (2001) Quantifying biodiversity: procedures and pitfalls in the measurement and comparison of species richness. *Ecology Letters*, **4**, 379-391.

Guisan, A. & Zimmermann, N. E. (2000) Predictive habitat distribution models in ecology. *Ecological Modelling*, **135**, 147-186.

Hatley, C. L. & MacMahon, J. A. (1980) Spider community organization: seasonal variation and the role of vegetation architecture. *Environmental Entomology*, **9**, 632-639.

Hirzel, A. & Guisan, A. (2002) Which is the optimal sampling strategy for habitat suitability modelling. *Ecological Modelling*, **157**, 331-341.

Hirzel, A., Hausser, J., Chessel, D. & Perrin, N. (2002) Ecological-Niche Factor Analysis: how to compute habitat-suitability maps without absence data? *Ecology*, **83(7)**, 2027-2036.

Hortal, J. & Lobo, J. M. (2002) Una metodología para predecir la distribución espacial de la diversidad biológica. *Ecología,* **16**, 405-432 + 4 láminas en color.

ITGE. (1988) *Atlas Geocientífico y del Medio Natural de la Comunidad de Madrid. Serie: Medio Ambiente.* Instituto Tecnológico GeoMinero de España, Madrid.

Jiménez-Valverde, A. & Hortal, J. (2003) Las curvas de acumulación de especies y la necesidad de evaluar la calidad de los inventarios biológicos. *Revista Ibérica de Aracnología*, **8**, 151-161.

Jiménez-Valverde, A. & Lobo, J. M. (2005) Determining a combined sampling procedure for a reliable estimation of Araneidae and Thomisidae assemblages (Arachnida: Araneae). *Journal of Arachnology*, **33**, 33-42

Johnston, C. A. (1998) *Geographic Information Systems in Ecology.* Blackwell Science, Oxford.

Legendre, P. & Legrendre, L. (1998) *Numerical Ecology.* Elsevier, Ámsterdam.

Lobo, J. M. (1997) Influencias geográficas, históricas y filogenéticas sobre la diversidad de las comunidades locales: una revisión y algunos ejemplos utilizando Scarabaeoidea coprófagos (Coleoptera, Laparosticti). *Boletín de la Asociación española de Entomología*, **21 (3-4)**, 15-31.

Lobo, J. M. (2000) ¿Es posible predecir la distribución geográfica de las especies basándonos en variables ambientales? En *Hacia un proyecto CYTED para el inventario y estimación de la diversidad entomológica en Iberoamérica: PrIBES 2000. m3m-Monografías Tercer Milenio, vol. 1*, eds. F. Martín-Piera, J. J. Morrone & A. Melic, pp. 55-68. Sociedad Entomológica Aragonesa (SEA), Zaragoza.

Martín, J. & Gurrea, P. (1999) Áreas de especiación en España y Portugal. Boletín de la *Asociación española de Entomología*, **23(1-2)**, 83-103.

Martín-Piera, F. & Lobo, J. M. (2000) Diagnóstico sobre el conocimiento sistemático y biogeográfico de tres órdenes de insectos hiperdiversos en España: Coleoptera, Hymenoptera y Lepidoptera. En *Hacia un proyecto CYTED para el inventario y estimación de la diversidad entomológica en Iberoamérica: PrIBES 2000*, eds. F. Martín-Piera, J. J. Morrone & A. Melic, pp. 287-308. Sociedad Entomológica Aragonesa (SEA), Zaragoza.

Martínez, J. A., Serrano, D. & Zuberogoitia, I. (2003) Predictive models of habitat preferences for the Eurasian eagle owl *Bubo bubo*: a multiscale approach. *Ecography*, **26**, 21-28.

Melic, A. (2000) Arañas de Aragón (Arácnida: Araneae). *Catalogus de la Entomofauna Aragonesa*, **22**, 3-40.

Melic, A. (2001) Arañas endémicas de la Península Ibérica e Islas Baleares (Arachnida: Araneae). *Revista Ibérica de Aracnología*, **4**, 35-92.

Myers, N. (2003) Conservation of Biodiversity: How are we doing? *The Environmentalist*, **23**, 9-15.

Ministerio de Agricultura, Pesca Y Alimentación (1986) *Atlas Agroclimático Nacional de España.* Dirección General de la Producción Agraria, Subdirección General de la Producción Vegetal, Madrid.

Mohler, C. L. (1983) Effect of sampling pattern on estimation of species distribution along gradients. *Vegetation*, **54**, 97-102.

New, T. R. (1999) Untangling the web: spiders and the challenges of invertebrate conservation. *Journal of Insect Conservation*, **3**, 251-256.

Peterson, A. T. & Kluza, D. A. (2003) New distributional modelling approaches for gap analysis. *Animal Conservation,* **6**, 47-54.

Peterson, A. T., Egbert, S. L., Sánchez-Cordero, V. & Price, K. P. (2000) Geographic analysis of conservation priority: endemic birds and mammals in Veracruz, Mexico. *Biological Conservation*, **93**, 85-94.

Reddy, S. & Dávalos, L. M. (2003) Geographical sampling bias and its implications for conservation priorities in Africa. *Journal of Biogeography*, **30**, 1719-1727.

Robinson, J. V. (1981) The effect of architectural variation in habitat on a spider community: an experimental field study. *Ecology*, **62**, 73-80.

Rushton, S. P. & Eyre, M. D. (1992) Grassland spider habitats in north-east England. *Journal of Biogeography*, **19**, 99-108.

Rushton, S. P., Topping, C. J. & Eyre, M. D. (1987) The habitat preferences of grassland spiders as identified using detrended correspondence analysis (DECORANA). *Bulletin of the British Arachnological Society*, **7**, 165-170.

Rypstra, A. L. (1986) Web spiders in temperate and tropical forests: relative abundance and environmental correlates. *American Midland Naturalist*, **115(1)**, 42-51.

Schadt, S., Revilla, E., Wiegand, T., Knauer, F., Kaczensky, P., Breitenmoser, U., Bufka, L., Červený, J., Koubek, P., Huber, T., Staniša, C. & Trepl, L. (2002) Assessing the suitability of central European landscapes for the reintroduction of Eurasian lynx. *Journal of Applied Ecology*, **39**, 189-203.

Soberón, J. & Llorente, B. J. (1993) The use of species accumulation functions for the prediction of species richness. *Conservation Biology*, **7**, 480-488.

Southwood, T. R. E. & Henderson, P. A. (2000) *Ecological Methods.* Blackwell Science.

StatSoft. (2001) *STATISTICA (data analysis software system and user´s manual). Version 6.* StatSoft, Inc., Tulsa, OK.

Store, R. & Jokimäki, J. (2003) A GIS-based multi-scale approach to habitat suitability modeling. *Ecological Modelling*, **169**, 1-15.

Suárez-Seoane, S., Osborne, P. E. & Alonso, J. C. (2002) Large-scale habitat selection by agricultural steppe birds in Spain: identifying species-habitat responses using generalized additive models. *Journal of Applied Ecology*, **39**, 755-771.

Toti, D. S., Coyle, F. A. & Miller, J. A. (2000) A structured inventory of Appalachian grass bald and heath bald spiders assemblages and a test of species richness estimator performance. *Journal of Arachnology*, **28**, 329-345.

Uetz, G. W. (1991) Habitat structure and spider foraging. En *Habitat structure: The Physical Arrangement of Objects in Space*, eds. S. S. Bell, E. D. McCoy & H. R. Mushinsky, pp. 325-348. Chapman and Hall, London.

Urones, C. & Majadas, A. (2002) Distribución espacial de Araneae según la arquitectura interna de los piornales montanos (*Cytisus oromediterraneus*). *Boletín de la Asociación española de Entomología*, **26(3-4)**, 93-105.

Urones, C. & Puerto, A. (1988) Ecological study of the Clubionoidea and Thomisoidea (Araneae) in the Spanish Central System. *Revue Arachnologique*, **8(1)**, 1-32.

Valdecasas, A. G. & Camacho, A. I. (2003) Conservation to the rescue of taxonomy. *Biodiversity and Conservation*, **12**, 1113-1117.

Wang, H. G., Owen, R. D., Sánchez-Hernández, C. & Romero-Almaraz, M. L. (2003) Ecological characterization of bat species distributions in Michoacán, México, using a geographic information system. *Global Ecology & Biogeography*, **12**, 65-85.

Wessels, K. J., Van Jaarsveld, A. S., Grimbeek, J. D. & Van der Linde, M. J. (1998) An evaluation of the gradsect biological survey method. *Biodiversity and Conservation*, **7**, 1093-1121.

Wiens, J. A., Addicott, J. F., Case, T. J. & Diamond, J. (1986) Overview: the importance of spatial and temporal scale in ecological investigations. En *Community Ecology,* eds. J. Diamond & T. J. Case, pp. 145-153. Harper and Row, Nueva York.

Williams, P. H., Margules, C. R. & Hilbert, D. W. (2002) Data requirements and data sources for biodiversity priority area selection. *Journal of Bioscience*, **27(Suppl. 2),** 327-338.

Wilson, E. O. (1999) *The Diversity of Live*. Norton, Nueva York.

Wise, D. H. (1993) *Spiders in Ecological Webs*. Cambridge University Press, Nueva York.

Zaniewski, A. E., Lehmann, A. & Overton, J. M. (2002) Predicting species spatial distributions using presence-only data: a case study of native New Zealand ferns. *Ecological Modelling*, **157**, 261-280.

# FACTORES DETERMINANTES DE LA RIQUEZA LOCAL DE ARAÑAS (ARANEIDAE & THOMISIDAE) A ESCALA REGIONAL: CLIMA Y ALTITUD VS. ESTRUCTURA DEL HÁBITAT

**RESUMEN.** El presente estudio analiza el efecto de factores locales sobre la riqueza regional de arañas (Araneidae y Thomisidae). Se obtuvieron inventarios fiables de 15 unidades territoriales de 1 $km^2$. La riqueza de especies se modelizó empleando Modelos Generales de Regresión y un conjunto de variables climáticas, topográficas y de estructura de la vegetación. Los efectos puros y conjuntos de cada grupo de factores se calcularon particionando la varianza explicada. Los resultados muestran la gran importancia de la complejidad de la vegetación, especialmente de los estratos herbáceo y subarbustivo, a la hora de determinar la riqueza de araneidos y tomísidos. La temperatura máxima es la única variable climática relacionada significativamente con la riqueza específica, aunque su efecto no puede separarse ni de otras variables con estructura espacial ni de la complejidad estructural del hábitat. Estos resultados apoyan la hipótesis de la heterogeneidad del hábitat y resaltan la importancia de considerar la complejidad de la vegetación a la hora de aplicar técnicas de manejo y conservar la fauna de arañas.

**Palabras clave:** Araneae, Araneidae, Thomisidae, Modelos Generales de Regresión, partición de la varianza, riqueza específica, estructura de la vegetación, Madrid, España

**Este capítulo ha sido enviado a publicar como:**

JIMÉNEZ-VALVERDE, A. & LOBO, J. M. Determinants of local spider (*Araneidae* & *Thomisidae*) species richness at a regional scale: climate and altitude vs. habitat structure. *Ecological Entomology*.

# DETERMINANTS OF LOCAL SPIDER (ARANEIDAE & THOMISIDAE) SPECIES RICHNESS ON A REGIONAL SCALE: CLIMATE AND ALTITUDE VS. HABITAT STRUCTURE

**ABSTRACT.** The present study analyzes the effect of local factors on regional spider (Araneidae & Thomisidae) richness. Fifteen territorial units of 1 $km^2$ were sampled to obtain reliable inventories of the two spider families. Richness values were modelled using General Regression Models and a set of climate, topographic and vegetation structure variables. Pure and joint effects were computed via variation partitioning. The results highlight the great importance of vegetation complexity, especially of grass and sub-shrub cover, in determining spider species richness. Maximum temperature is the only climate variable significantly related with species richness, although its effect is combined with that of spatial and vegetation structure variables. These results support the habitat heterogeneity hypothesis and highlight the importance of taking vegetation complexity into account when managing habitats and spider conservation is desired.

**Key words:** Araneae, Araneidae, Thomisidae, General Regression Models, variation partitioning, species richness, vegetation structure, Madrid, Spain

## INTRODUCTION

About 40000 spider species are presently known (Platnick, 2005), although estimations of their total number vary from 60000 to 170000 (Coddington & Levi, 1991). Generalist predators colonizing almost all habitats, quite abundant and diverse in

natural systems (density values around 150 individuals/$m^2$, Nyffeler, 2000), they develop a great variety of life histories, behaviour, and morphologic, physiological and ecological adaptations (Turnbull, 1973; Wise, 1993; Foelix, 1996). Because of their diverse relationships with the environment and their impact on prey populations (Nyffeler, 2000), spiders have been proposed as very suitable for pest limitation and bioindication (Clausen, 1986; Marc *et al*., 1999). Araneidae, one of the most successful spider families (approximately 2600 species; Foelix, 1996), are relatively easy to detect due to their size, coloration, and their orb webs. Unlike the araneids, Thomisidae (crab spiders) do not use webs to capture prey, instead ambushing prey from flowers or leaves, where their cryptic coloration allows them to go unnoticed. Some genera, like *Xysticus* and *Ozyptila*, live primarily among leaf litter and herbaceous vegetation.

Habitat structure and, more precisely, vegetation complexity, has been consistently recognized as one of the most important factors in determining the presence of spider species, as well as their species richness and composition (Colebourn, 1974; Hatley & Macmahon, 1980; Robinson, 1981; Urones & Puerto, 1988; Döbel *et al*., 1990; Uetz, 1991; Wise, 1993; Downie *et al*., 1995; Balfour & Rypstra, 1998; Downie *et al.*, 2000; Borges & Brown, 2001). Thus, despite the absence of strong spider association with host-plants, vegetation type can be an important factor in determining spider assemblages due to their relationship with vegetation structure (see Urones & Puerto, 1988). Additionally, other climate and topographic factors have been pointed out as relevant for spiders: humidity (Coulson & Butterfield, 1986; Rushton *et al*., 1987; Rushton & Eyre, 1992; Bonte *et al*., 2002), temperature (Rypstra, 1986), and altitude (Urones & Puerto, 1988; Rushton & Eyre, 1992; Chatzaki *et al*., 2005). However, the relative importance of these factors has scarcely been investigated. Greenstone (1984) noticed that web-building spider diversity along altitude transects in both tropical and

temperate localities was mainly determined by vegetation structure, while elevation and climate differences between localities had no important consequences. On the other hand, Rypstra (1986) measured web-building spider richness in temperate, subtropical and tropical localities within an area of 3 hectares, stating that spider diversity was also mainly determined by the vegetation structure of each locality, followed by prey availability and environment temperature, while relative humidity had no effect.

The main problem with these studies, based on correlations, is the unavoidable collinearity of explanatory variables, which limit regression analysis adequacy in finding appropriate causal variables (Mac Nally, 2000). As direct causal relationships are never known because of the lack of detailed physiological and autoecological studies, variables widely considered to be influential on species diversity and introduced in regression analysis are not necessarily causally related, even though correlated, with species richness (Legendre & Legendre, 1998; Mac Nally, 2000). Results in search of causality are worsened by collinearity affecting automatic variable selection methods. Additionally, although statisticians emphasize that a model without proper validation has no merit (Olden & Jackson, 2000), no validation of models is reported in any of these two formerly cited studies.

In order to preserve spider biodiversity, land management strategy design requires an understanding of patterns of spider diversity on the appropriate regional scale (New, 1999). Thus, in an effort to overcome the formerly mentioned drawbacks, the effect of local factors on regional spider (Araneidae & Thomisidae) richness has been analysed using variation partitioning techniques (Borcard *et al*., 1992) to estimate the predictive capacity of each explanatory variable.

## METHODS

**The area of study.—** The Comunidad de Madrid, with approximately 8028 $km^2$, is located in Central Iberia (Fig. 1). Although mean altitude is around 800 m.a.s.l., it ranges from 2430 m.a.s.l. in the ‘‘Sistema Central’’ mountain range to 434 m.a.s.l. in the Alberche valley. Its heterogeneous lithology includes acid-rock mountains (granite and gneiss), a ramp of acidic and coarse-grained sands, many alluvial, fine-grained soils in lowlands, and a clay, limestone and gypsum soil plateau. The region has a continental climate with Mediterranean influence and annual precipitations ranging from 350 mm to 2000 mm. The phytosociological characteristics of the Comunidad de Madrid are its location in the Mediterranean region and its hosting of representatives of the Mesomediterranean, Supramediterranean, Oromediterranean and Crioromediterranean flora (Rivas-Martínez, 1987).

**Selection of the sampling territorial units. —** Geographic biodiversity pattern description requires reliable inventories which recover the maximum variation range of the focus taxa in the selected territory, so a good design of sampling locations is essential. We have used the methodology proposed by Jiménez-Valverde & Lobo (2004), based on cluster analysis, designed to regionalize a territory and to maximize field work accomplished, while considering both the variables that a priori most affect the focus taxa and the spatial location of the sampling points. Sampling effort is previously defined, being the initial realistic criterion which determines the number of sampling points to chose. Thus, this method maximizes the spatial-environmental variation recovered as a function of sampling effort (see Jiménez-Valverde & Lobo, 2004 for a detailed discussion of the method).

Using a Geographic Information System (Idrisi 32, Clark Labs, 2000a), five environmental layers were created, namely vegetation structure, precipitation,

temperature, altitude and lithology, as these are widely recognized factors affecting spider distribution (see Introduction). The Corine Land Cover map (European Environment Agency, 1996) was reclassified in three broad categories representing structural complexity: grasslands, scrublands and forests. Mean annual temperature and total annual precipitation were courtesy of the Spanish Instituto Nacional de Meteorología. Mean altitude was obtained from a global digital elevation model (Clark Labs, 2000b). A lithologic map (ITGE, 1988) was digitized and reclassified into basic and acidic soils. Those sampling territorial units with a dominance of urban or agricultural land use, of bare rock, bodies of water or sand were not included. These five environment variables, together with two spatial variables (central latitude and longitude of each 1 $km^2$ sampling territorial unit) were used to divide the Comunidad de Madrid territory into 15 subregions by a *k-means* cluster analysis, maximizing the initial distance among the initial centroids (StatSoft, 2001). To select the final 15 sampling territory units (Fig. 1), two hierarchic criteria were followed: 1) distance to the cluster centroid, such that the smaller the distance, the better the spatio-environmental representation of the cluster; and 2) ease of access.

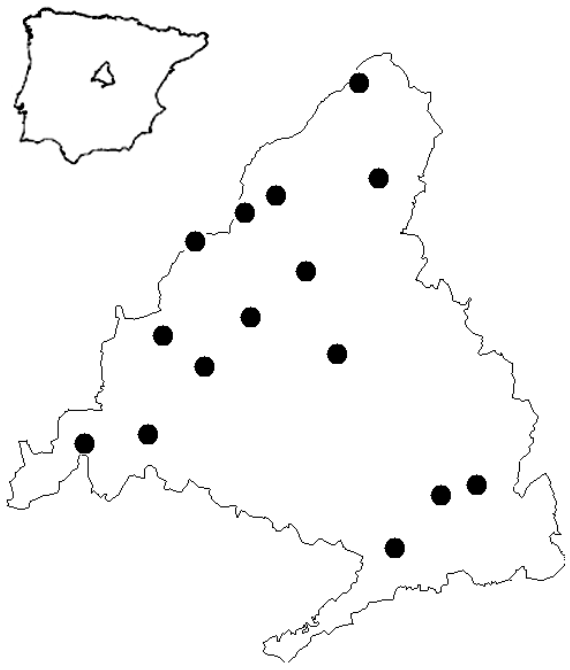


Figure 1.- Location of the 1 $km^2$ sampling plots in Madrid, central Iberian peninsula. The selection of the plots by a *k-means* procedure to maximize the environmental and spatial variation of the region (see Jiménez-Valverde & Lobo, 2004).

**Sampling method. —** The sampling protocol developed by Jiménez-Valverde & Lobo (2005) was employed, as it yields reliable inventories of both families (Araneidae and Thomisidae) in 1 $km^2$ sampling plots. Each 1 $km^2$ sampling territorial unit (=sampling plot) was divided into 2500 subplots of 400 $m^2$; 20 of these subplots were chosen at random, and a sampling effort unit carried out in each. A sampling effort unit was defined as the combined effort of the following sampling methods: 1) A one-person sweep of the herbaceous vegetation and shrub during 15 minutes. 2) A one-person beating of bushes and small trees and branches during 15 minutes with a heavy stick; the specimens fell on a 1.25 × 1.25 m white sheet. In cases where the structure of the vegetation made the use of the sheet difficult a 41 × 29 cm plastic pail was employed. 3) The running of 4 open pitfall traps during 48 hours. These traps were 11.5 cm wide and 1 liter in volume, each 10 m apart to avoid interference effects and to maximize their efficacy. Traps were filled with water, and a few drops of detergent added to break the surface tension so as to prevent the spiders from escaping. Additionally, in places where, due to special habitat structure, araneids tend to concentrate in particular habitat patches, a one-person visual search from knee level to as high as one can reach during 15 minutes was also added to the sampling protocol (see Jiménez-Valverde & Lobo, 2005 for a detailed description of the protocol).

**Sampling-related considerations. —** Juveniles usually are discarded in spider biodiversity studies (e. g. Jerardino *et al*., 1991; Toti *et al*., 2000; Sørensen *et al*., 2002) because of their difficult identification (Coddington *et al*., 1996; Dobyns, 1997). However, their inclusion is necessary to obtain reliable estimates of species richness (Jiménez-Valderde & Lobo, 2006). Thus, in this study juveniles that could be identified to the species level were included in the analysis. This was possible for many araneid and some thomisid species which have a distinguishing, characteristic color pattern, and

for some genera represented by only one species in the Iberian Peninsula (i.e., *Mangora acalypha* (Walckenaer, 1802), *Runcinia grammica* (Koch, C.L., 1837) and *Synaema globosum* (Fabricius, 1775)).

Unfortunately, long-term intensive sampling is not often affordable in biodiversity surveys, especially those with multiple sampling points. So, in each sampling plot we performed an exhaustive sampling protocol as detailed above in spring. This strategy yields reasonable estimates of the entire spring fauna, enabling sites to be effectively compared, since spring inventories are a relatively good representation both of the annual species richness and faunistic composition (Jiménez-Valverde & Lobo, 2006). Moreover, short intensive samplings avoid the effects of immigration, thus enabling more robust comparative analysis (Samu & Lövei, 1995; Sorensen, 2004).

***In situ* structural variables. —** Vegetation complexity was measured in each sampling locality by employing a modified version of the method developed by Newsome & Catling (1979) (see also Coops & Catling, 1997a, b, 2000). Vegetation structure was visually assessed in each subsampling plot using seven habitat features (tree canopy cover, TREE; shrub canopy cover, SHRUB; sub-shrub canopy cover, SSHRUB; ground herb cover, GRASS; soil moisture, SMOIST; amount of leaf litter, LITTER; amount of logs, rocks and debris, LRD), scored between 0 and 3 using ordinal scales (see Table 1). Then, a mean score was calculated for each habitat feature in each 1 $km^2$ sampling unit and a global score for vegetation complexity (VEG) was obtained by adding the 7 partial values. From these 7 values for each sampling unit, just TREE and LITTER were significantly correlated ($r$=0.92, $p < 0.05$). This method has proved to be useful for predicting variations in species richness and composition of ants (Lassau & Hochuli, 2004), wasps (Lassau & Hochuli, 2005) and beetles (Lassau *et al.*, 2005a).

Table 1.- Visual scoring of vegetation structure features in each subsampling plot. *Sparse ground flora: grasses covering <50% of the subsampling plot. **Dense ground flora: grasses covering >50% of the subsampling plot (adapted from Coops & Catling, 1997b).

| Structure | Score | | | |
|---|---|---|---|---|
| | **0** | **1** | **2** | **3** |
| **Tree canopy cover** | 0 | <30 | 30-70 | >70 |
| **Shrub canopy cover** | 0 | <30 | 30-70 | >70 |
| **Sub-shrub canopy cover** | 0 | <30 | 30-70 | >70 |
| **Ground flora** | Sparse* (<0.5) | Sparse* (>0.5) | Dense** (<0.5) | Dense** (>0.5) |
| **Logs, rocks and debris** | 0 | <30 | 30-70 | >70 |
| **Soil moisture** | Dry | Moist | Permanent water adjacent | Water-logged |
| **Leaf litter** | 0 | <30 | 30-70 | >70 |

**Statistical analyses. —** *The dependent variable.* As inventories are almost always incomplete, and the degree of incompleteness varies, it is necessary to reduce survey bias to work with data as accurate as possible (Hortal, 2004; Hortal *et al*., 2004). For each location, species accumulation curves were drawn (see Soberón & Llorente, 1993; Gotelli & Colwell, 2001; Jiménez-Valverde & Hortal, 2003) using subsampling plots as sampling effort units (Fig. 2). The order in which sampling effort units were added was randomized 500 times to build smoothed curves using EstimateS 7.5 software (Colwell, 2005). As completeness of inventories differed from location to location, no single non-parametric estimator performed well in all 15 sampling plots; the six calculated (Chao1 and 2, Jackknife1 and 2, ICE and ACE) produced nonsense estimations for different sampling plots (see Brose *et al*., 2003). Hortal *et al.* (2004), using the Clench estimation as dependent variable in their analysis of butterfly species richness in mainland Portugal, produced more consistent models with fewer residual errors in the prediction than by using observed species richness directly. Therefore, we decided to estimate the asymptotic value of the accumulation curves using the Clench

equation (Soberón & Llorente, 1993; Colwell & Coddington, 1994; León-Cortés *et al*., 1998; Peterson & Slade, 1998). The models were fitted to the data through non-linear regression using the Simplex & Quasi-Newton algorithm (StatSoft, 2001). Although other more complex models have been recommended (Flather, 1996; Jiménez-Valverde *et al*., 2006), nonsense predicted species richness values discouraged their use (Jiménez-Valverde *et al*., 2006). The estimated species richness values of the Clench model were considered as the dependent variable in subsequent analysis.

*The modelling process.* Multiple relationships between species richness and the explanatory variables were analysed using general regression models (GRM) (StatSoft, 2001). The continuous variables selected are: mean altitude (ALT), annual precipitation (PRECP), precipitation of the least rainy month (precipitation of August, PAUG), precipitation of the most rainy month (precipitation of April, PAP), mean annual temperature (TEMP), maximum and minimum annual temperature (TMAX and TMIN respectively), insolation (INS), lithology (basic and acid soils, LIT), land use following Corine (reclassified in grasslands, shrubland and forests, LUU), and vegetation structure scores (partial and global) (all climate variables were provided by the Spanish Instituto Nacional de Meteorología). Continuous variables were standardized to 0 mean and 1 standard deviation to avoid scale effects. As a first step, categorical variables and linear and quadratic functions of the continuous variables were regressed independently against the response variable to determine significant predictors. Then, significant terms were sequentially introduced in the model according to their change in deviance and selected by a backward stepwise procedure. Spatial variables were included in the model after environmental variables to account for effects caused by other unaccounted-for historic, biotic or environment variables and also to eliminate the probable spatial autocorrelation in the residuals (Legendre & Legendre, 1998). The terms of the third-

degree polynomial equation of the central latitude (LAT) and longitude (LON) of each square ($ß_1LAT + ß_2LON + ß_3LAT^2 + ß_4LAT \cdot LON + ß_5LON^2 + ß_6LAT^3 + ß_7LAT^2 \cdot LON + ß_8LAT \cdot LON^2 + ß_9LON^3$) were independently tested for significance. The significant spatial terms were added to the model and subjected to a backward stepwise procedure together with environmental predictors.

The reliability of the final model was checked using a Jackknife procedure, in which 15 models where recalculated leaving out one sampling plot in turn, then calculating the estimated species richness for each plot (see Hortal *et al*., 2001). The predictive error (ME) of the final model was calculated as the mean of:

$$E_i=(|O_i-P_i|)/O_i \times 100$$

where $O_i$ is the observed species in each plot and $P_i$ is the predicted value (Pascual & Iribarne, 1993), being 100 - ME the predictive power of the model. Outliers were identified as those observations in which the residual absolute value is greater than the standard deviation of the predicted values (Nicholls, 1989). Outliers were checked to determine if they were due to erroneous data or to unique environmental conditions; while the former must be discarded, the latter should be included in the analysis to account for such special environmental conditions (Hortal *et al*., 2001). After removing outliers, models were fitted again.

Finally, variation partitioning of significant explanatory variables was used to quantify the relative importance of the effect of each determinant alone, and its respective shared influences (Legendre & Legendre, 1998). After rejecting those variables not related with the dependent variable, total variation was decomposed among three groups of variables (the most important ones): EV = environment variables (maximum annual temperature), SV = spatial variables (latitude) and VEGS = vegetation structure variables (sub-shrub cover, grassland cover), and the percentage of

explained deviance calculated for eight different components: a = sole effect of EV alone, b = sole effect of SV alone, c = sole effect of VEGS alone, d = combined variation due to the joint effect of EV and SV, e = combined variation due to the joint effect of EV and VEGS, f = combined variation due to the joint effect of SV and VEGS, g = combined variation due to the joint effect of the three components, and variation not explained by the independent variables included in the analysis (U). The decomposition of the variation in species richness into the three sets of explanatory variables was carried out by means of a partial regression analysis (Legendre & Legendre, 1998). Such an approach allows one to deal with dependent explanatory variables, as it is explicitly designed to identify the portions of explained variability that are shared by different factors, and those that are independent (Heikkinen *et al*., 2004; Lobo *et al*., 2004). In the process of variation decomposition, species richness (*y*) was regressed with the three types of variables together (EV, SV and VEGS), which represent the total explained variation in the data set ($a + b + c + d + e + f + g$ in Fig. 1). Regressing *y* with each one of the explanatory variables yields the variation separately attributable to EV ($a + d + e + g$), SV ($b + d + f + g$), and VEGS ($c + e + f + g$). Subsequently, residuals of the regression of EV against SV + VEGS variables were calculated, and *y* was regressed with these residuals to estimate the sole effect of EV variation (*a*). Fractions *b* and *c* were estimated in the same way after computing the regression residuals of SV against EV + VEGS, and the regression residuals of VEGS against EV + SV, respectively. The remaining variation fractions were computed according to two sets of equations (Borcard *et al.*, 1992), where:

$$d + e + g = \text{EV} - a$$

$$d + f + g = \text{SV} - b$$

$$e + f + g = \text{VEGS} - c$$

and

$$d = (EV + SV) - (e + f + g) - (a + b)$$

$$e = (EV + VEGS) - (d + f + g) - (a + c)$$

$$f = (SV + VEGS) - (d + e + g) - (b + c)$$

$$g = (d + e + g) - d - e = (d + f + g) - d - f = (e + f + g) - e - f$$

## RESULTS

Accumulation curves of the 15 inventories indicate quite complete inventories, almost all approaching the asymptote (Fig. 2). The most notable exceptions were La Herreria and Cerro Cardoso (dotted and thick line in Fig. 2, respectively), which end while still rising. The latter is specially striking as it rises quite slowly, indicating an extremely slow addition of the species to the inventory. Clench estimations indicate that between 71% and 92 % of the fauna have been collected, except in Cerro Cardoso where the estimation of the function is quite high, yielding a value of completeness of 47 %.

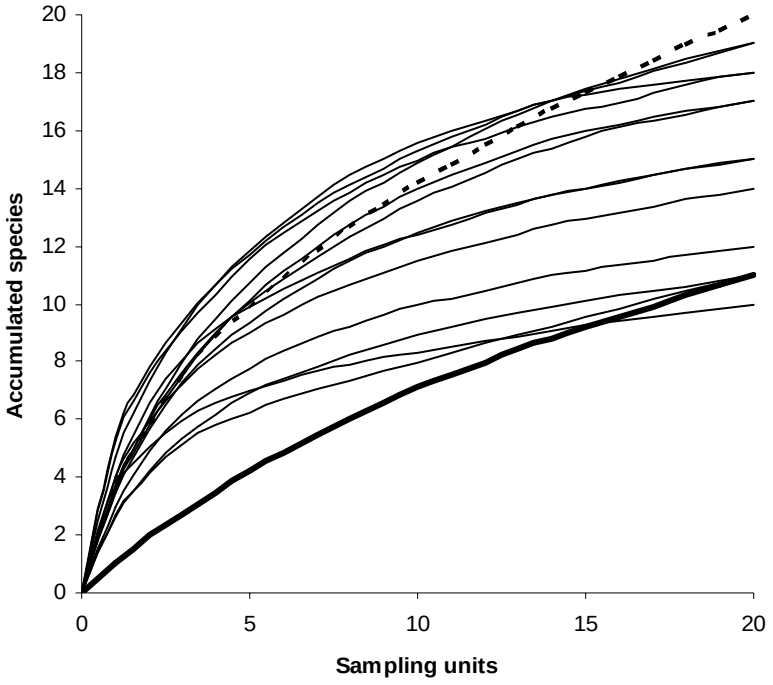



Figure 2.- Accumulation curves for the 15 inventories (thick line, Cerro Cardoso; dotted line, Perales de Tajuña).

Only two of the 10 environment variables tested were significantly related with species richness, TMAX (as a quadratic function, see Fig. 3) and LIT (Table 2A). Only TMAX remains when submitted with LIT to a backward stepwise selection procedure. Of the 8 *in situ* structural variables, only VEG, GRASS and SSHRUB accounted for a significant explained variance, showing a positive relationship with species richness, and only the latter two remained in a complete model with structural variables. LONG (as a quadratic function) and LAT were the only spatial variables significantly related with the dependent variable, but only LAT remains after a backward stepwise selection. The final whole model built with all formerly significant variables only retained GRASS and SSHRUB, accounting for 61% of total variability in species richness. The jackknife procedure yields a predictive power of 86.2%, and Cerro Cardoso and Perales de Tajuña were identified as possible outliers. As the first sampling plot was the one with the greatest absolute residual value, it was analyzed first. It was recognized as a true outlier, because the addition of species to the inventory was so slow (see Fig. 2) being the estimation of the Clench function highly unreliable. So Cerro Cardoso was eliminated from the matrix and the models fitted again.

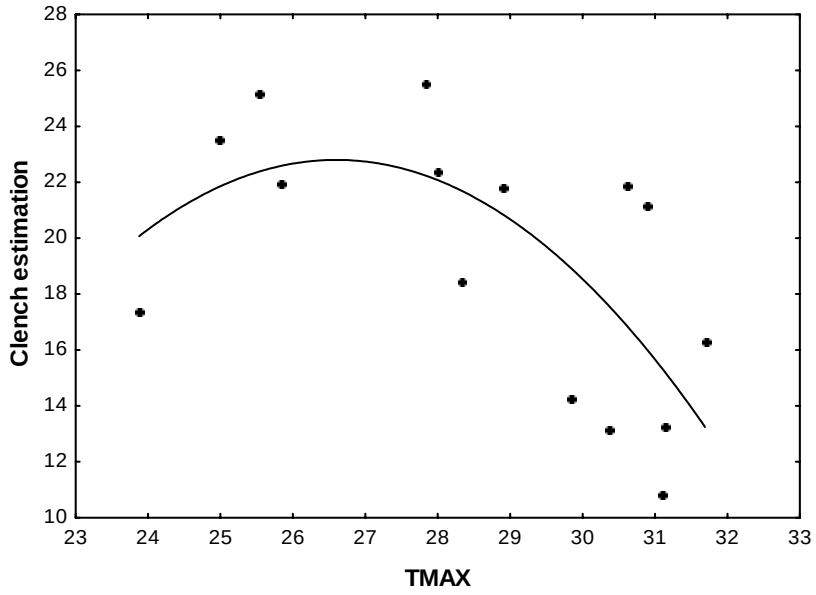



Figure 3.- Relationship between species richness (Clench estimation) and maximum temperature (in ºC).

After removing the outlier, the variables significantly related with species richness are the same as before except LONG, which no longer remained (Table 2B). Variance explained varied little, with GRASS and SSHRUB the variables which most increased. In fact, the whole model retains only these two structural variables explaining 81% of the variability of the dependent variable. The predictive power increased to 88.6%, and no outlier was identified.

Table 2.- Statistiscally significant variables related to spider species richness and models for each group of variables, respective regression coefficients and percentage of explained variance ($R^2$ (%)). $f^2$ is the quadratic function of the variable considered (A, with the 15 sampling plots; B, deleting the outlier of Cerro Cardoso).

| Variable | Function | $R^2$ (%) | *F* | *p* |
|---|---|---|---|---|
| **A** | | | | |
| Tmax | Quadratic | 36.40 | 5.01 | 0.026 |
| Lit | - | 34.58 | 8.40 | 0.012 |
| Veg | Linear (+) | 29.50 | 6.86 | 0.021 |
| Grass | Linear (+) | 35.70 | 8.77 | 0.011 |
| Sshrub | Linear (+) | 30.62 | 7.18 | 0.019 |
| Long | Quadratic | 32.70 | 4.40 | 0.037 |
| Lat | Linear (+) | 34.65 | 8.42 | 0.012 |
| Model for EV | $f^2$ Tmax | 36.40 | 5.01 | 0.026 |
| Model for SV | Lat | 34.65 | 8.42 | 0.012 |
| Model for VEGS | Grass+Sshrub | 61.19 | 12.04 | 0.001 |
| Whole model | Grass+Sshrub | 61.19 | 12.04 | 0.001 |
| **B** | | | | |
| Tmax | Quadratic | 32.45 | 4.12 | 0.046 |
| Lit | - | 31.09 | 7.43 | 0.018 |
| Veg | Linear (+) | 26.92 | 5.79 | 0.033 |
| Grass | Linear (+) | 42.48 | 10.60 | 0.007 |
| Sshrub | Linear (+) | 42.84 | 10.74 | 0.007 |
| Lat | Linear (+) | 30.49 | 6.70 | 0.024 |
| Model for EV | $f^2$ Tmax | 32.45 | 4.12 | 0.046 |
| Model for SV | Lat | 30.49 | 6.70 | 0.024 |
| Model for VEGS | Grass+Sshrub | 81.03 | 28.76 | <0.001 |
| Whole model | Grass+Sshrub | 81.03 | 28.76 | <0.001 |

A considerable proportion of the variation in species richness is due to the joint effect of the three types of variables considered, showing that around a 30% of total variability in species richness is not attributable to a single variable. However, variation partitioning highlights the important effect of vegetation structure alone (57%), and the lack of relevance of environment and spatial variables by themselves (Fig. 4). Both environment and spatial variables are important due to their joint effect (8.9%), indicating that spatially structured environmental factors seem to influence the variation in species richness slightly. Lastly, the negative signs of the interaction between vegetation structure with environment or spatial variables (-7.0% and -8.9%, respectively) suggests the probably moderate synergic effects of these variables (see Legendre & Legendre, 1998).

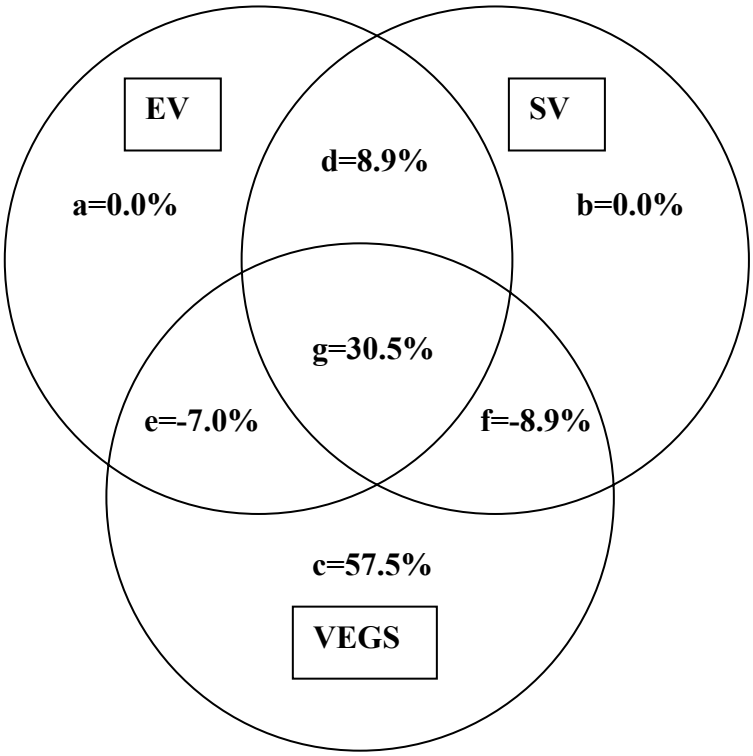



Figure 4.- Variation partitioning in species richness between the statistically significant variables (EV = environment variables (maximum annual temperature), SV = spatial variables (latitude) and VEGS = vegetation structure variables (sub-shrub cover, grassland cover)).

## DISCUSSION

The habitat heterogeneity hypothesis states that the more complex the habitat, the more niches available, and therefore the higher the species richness (Tews *et al*., 2004). The bias (focus on vertebrates and on habitats under anthropogenic pressure) of studies dealing with this topic, together with the limited variety of spatial and temporal scales on which the effects of vegetation structure have been tested, impede comparisons and conclusive general results (McCoy & Bell, 1991; Tews *et al*., 2004). Even though studies on spiders are not free of temporal and spatial scale problems, still habitat complexity has been repeatedly indicated as the most important factor in determining spider distribution. In this study, by employing a measure of habitat complexity in accordance with our spatial scale of analysis, vegetation complexity appears as a powerful predictor of local spider species richness on a regional scale. Of all the structural variables tested, only herbaceous and sub-shrub cover remains in the final model as significant; their sole effect is considerable, reflecting the main range of height occupied by Araneidae and Thomisidae in natural habitats. Grill *et al*. (2005) reported that the cover of herbaceous vegetation is the main determining factor for web-building spider diversity on a local scale in Mediterranean habitats of Sardinia. As thomisids live mainly on vegetation, with a great percentage of species living in the soil-grass interface (genus *Xysticus* and *Ozyptilla*), this variable is quite relevant for both families.

The model developed would be more useful if it could be extrapolated to the entire region to map richness predictions to be used in conservation planning. However, because *in situ* structural variables are not available for the whole territory, such representation is impossible. Nevertheless, several authors have pointed out that the

complexity measures used in this work can be recognized in airborne video graphic data (Coops & Catling, 1997a, b). Moreover, Lassau *et al*. (2005b) found a good correlation between Normalized Difference Vegetation Index (NDVI) scores and these structural measures. NDVI provides a means of monitoring density and vigour of green vegetation growth since Geographic Information System layers are available from a number of organizations (Pettorelli *et al*., 2005); thus, it is a widely used factor in studies of species-richness patterns and in predictive modelling (i.e., Egbert *et al*., 2002; Suárez-Seoane *et al*., 2002; Bailey *et al*., 2004; Foody, 2004; Parra *et al*., 2004; Roura-Pascual *et al*., 2004; Ruggiero & Kitzberger, 2004). However, we tested correlations between our structural variables and NDVI scores from a layer of 1 $km^2$ resolution, calculated as the mean maximum monthly value for April, May and June of the year 2001 (courtesy of the CREPAD, Instituto Nacional de Técnica Aeroespacial, Gran Canaria, Spain), and found none statistically significant. Moreover, although NDVI was significantly related with species richness, it did not remain in any model, neither in the environmental nor in the final one. The lack of concordance between our results and those from Lassau *et al*. (2005b) may be due to differences in resolution, and perhaps our greater plot size dilutes the relationship between both variables.

Although vegetation structure is widely recognized as one of the main determinants of spider community composition, the exact mechanism of its influence is unknown (Wise, 1993; Rypstra *et al.*, 1999). Availability of structures for attaching the web, ambush and refuge sites is probably the most direct effect of vegetation complexity, but other indirect effects may be related, such as, for example, microclimate, prey availability or reduced cannibalism (Uetz, 1991; Marc *et al*. 1999). Many more field studies would be necessary to clarify this interesting and relevant question.

Maximum temperature seems to be a variable slightly related with species richness, although its effect is combined with those of spatial and vegetation structure variables. Species richness is maximized at intermediate values of this factor (~26.5 ºC), diminishing at low and high values (see Fig. 3). Negative interaction effects between vegetation structure variables and environment or spatial ones suggest that both pairs of factors are able to explain more than the sum of their individual effects: i.e. more species richness can be observed when the adequate vegetation structure occurs in a distinctive spatial location or throughout some characteristic climate conditions. It is interesting to note that low maximum temperatures generally occur in high altitudes and latitudes, and that the structure of the vegetation can acquire more relevance in this situation that in lowland or warm places.

Araneidae and Thomisidae are two families that have a high proportion of species with wide distributions. This fact may be related with their great dispersal potential, as Araneidae and Thomisidae are, after Linyphiidae, the two most numerous families of ballooning spiders (Dean & Sterling, 1985; Bishop, 1990). So, they are species with broad environment tolerance but with a great dependence on the physical structure of the environment due to their life histories. These facts highlight the importance of accounting for the preservation of vegetation complexity in management planning. Bell *et al*. (2001) reviewed how both intensity and type of management in grasslands and heathlands affect spider assemblages, and recommended preserving natural cover in conditions of intensive habitat pressure (grazing, cutting, etc.). Our results may probably be extrapolated to other epiphytic spider families such as, for example, Clubionidae or Philodromidae. On the contrary, ground-dwelling families (for example, Gnaphosidae or Lycosidae) may respond differently to environmental factors, and other variables such as humidity may constrain their diversity patterns (Grill *et al*.

2005). Additional research in this direction is necessary to make the reliable inclusion of spiders in management decisions possible.

## AKCNOWLEDGEMENTS

To the Consejería de Medio Ambiente y Ordenación del Territorio of the Comunidad de Madrid for the survey authorizations. To Ángel García Sevilla, from CREPAD, for his help with NDVI images. Comments of Joaquín Hortal greatly improved the manuscript. This paper has been supported by a Fundación BBVA project (Diseño de una red de reservas para la protección de la Biodiversidad en América del sur austral utilizando modelos predictivos de distribución con taxones hiperdiversos) and a MEC Project (CGL2004-04309), as well as by a Museo Nacional de Ciencias Naturales/C.S.I.C./Comunidad de Madrid PhD grant.

## LITERATURE CITED

Bailey, S.-A., Horner-Devine, M. C., Luck, G., Moore, L. A., Carney, K. M., Anderson, S., Betrus, C. & Fleishman, E. (2004) Primary productivity and species richness: relationship among functional guilds, residency groups and vagility classes at multiple spatial scales. *Ecography*, **27**, 207-217.

Balfour, R. A. & Rypstra, A. L. (1998) The influence of habitat structure on spider density in a no-till soybean agroecosystem. *Journal of Arachnology*, **26**, 221-226.

Bell, J. R., Wheater, C. P. & Cullen, W. R. (2001) The implications of grassland and heathland management for the conservation of spider communities: a review. *Journal of Zoology*, **255**, 377-387.

Bishop, L. (1990) Meteorological aspects of spiders balloning. *Environmental Entomology*, **19**, 1381-1387.

Bonte, D., Baert, L. & Maelfait, J.-P. (2002) Spider assemblage structure and stability in a heterogeneous coastal dune system (Belgium). *Journal of Arachnolgy*, **30**, 331-343.

Borcard, D., Legendre, P. & Drapeau, P. (1992) Partialling out the spatial component of ecological variation. *Ecology*, **73**, 1045-1055.

Borges, P. A. V. & Brown, V. K. (2001) Phytophagous insects and web-building spiders in relation to pasture vegetation complexity. *Ecography*, **24**, 68-82.

Brose, U., Martínez, N. D. & Williams, R. J. (2003) Estimating species richness: sensitivity to sample coverage and insensitivity to spatial patterns. *Ecology*, **84**, 2364-2377.

Chatzaki, M., Lymberakis, P., Markakis, G. & Mylonas, M. (2005) The distribution of ground spiders (Araneae, Gnaphosidae) along the altitudinal gradient of Crete, Greece: species richness, activity and altitudinal range. *Journal of Biogrography*, **32**, 813-831.

Clark Labs. (2000a) *Idrisi 32.02. GIS software package.* Clark University.

Clark Labs. (2000b) *Gobal Change Data Archive Vol. 3. 1 km Global Elevation Model.* CD-Rom. Clark University.

Clausen, I. H. S. (1986) The use of spiders as ecological indicators. *Bulletin of the British Arachnological Society*, **7(3)**, 83-86.

Coddington, J. A. & Levi, H. W. (1991) Systematics and evolution of spiders. *Annual Review of Ecology and Systematics*, **22**, 565-592.

Coddington, J. A., Young, L. H. & Coyle, F. A. (1996) Estimating spider species richness in a southern Appalachian cove hardwood forest. *Journal of Arachnology*, **24**, 111-128.

Colebourn, P. H. (1974) The influence of habitat structure on the distribution of *Araneus diadematus* Clerck. *Journal of Animal Ecology*, **43**, 401-409.

Colwell, R. K. & Coddington, J. A. (1994) Estimating terrestrial biodiversity through extrapolation. *Philosophical Transactions of the Royal Society (series B)*, **345**, 101-118.

Colwell, R. K. (2005*) EstimateS: Statistical Estimation of Species Richness and Shared Species from Samples (Software and User´s Guide), Version 7.5*, available in http://viceroy.eeb.uconn.edu/estimates

Coops, N. C. & Catling, P. C. (2000) Estimating forest habitat complexity in relation to time since fire. *Austral Ecology*, **25**, 344-351.

Coops, N. C. & Catling, P. C. (1997a) Predicting the complexity of habitat in forests from airborne videography for wildlife management. *International Journal of Remote Sensing*, **18(12)**, 2677-2682.

Coops, N. C. & Catling, P. C. (1997b) Utilising airborne multispectral videography to predict habitat complexity in eucalypt forests for wildlife management. *Wildlife Research*, **24**, 691-703.

Coulson, J. C. & Butterfield, J. (1986) The spider communities on peat and upland grasslands in northern England. *Holartic Ecology*, **9**, 229-239.

Dean, D. A. & Sterling, W. L. (1985) Size and phenology of ballooning spiders at two locations in eastern Texas. *Journal of Arachnology*, **13**, 111-120.

Döbel, H. G., Denno, R. F. & Coddington, J. A. (1990) Spider (Araneae) community structure in an intertidal salt marsh: effects of vegetation structure and tidal flooding. *Environmental Entomology*, **19**, 1356-1370.

Dobyns, J. R. (1997) Effects of sampling intensity on the collection of spider (Araneae) species and the estimation of species richness. *Environmental Entomology*, **26**, 150-162.

Downie, I. S., Butterfield, J. E. L. & Coulson, J. C. (1995) Habitat preferences of sub-montane spiders in northern England. *Ecography*, **18**, 51-61.

Downie, I. S., Ribera, I., McCracken, D. I., Wilson, W. L., Foster, G. N., Waterhouse, A., Abernethy, V. J. & Murphy, K. J. (2000) Modelling populations of *Erigone atra* and *E. dentipalpis* (Araneae: Linyphiidae) across an agricultural gradient in Scotland. *Agriculture, Ecosystems and Environment*, **80**, 15-28.

Egbert, S. L., Martínez-Meyer, E., Ortega-Huerta, M. & Peterson, A. T. (2002) *Use of datasets derived from timeseries AVHRR imagery as surrogates for land cover maps in predicting species' distributions.* Proc. IEEE–Int. Geoscience and Remote Sensing Symp., I–IV, pp. 2337–2339 New York: IEEE Publishers.

European Environment Agency. (1996) *Natural Resources CD-Rom.* European Environment Agency.

Flather, C. H. (1996) Fitting species-accumulation functions and assessing regional land use impacts on avian diversity. *Journal of Biogeography*, **23**, 155-168.

Foelix, R.F. (1996) *Biology of spiders*. Oxford University Press, New York.

Foody, G. M. (2004). Spatial nonstationary and scale-dependency in the relationship between species richness and environmental determinants for the sub-Saharan endemic avifauna. *Global Ecology and Biogeography*, **13**, 315-320.

Gotelli, N.J. & Collwell, R. K. (2001) Quantifying biodiversity: procedures and pitfalls in the measurement and comparison of species richness. *Ecology Letters*, **4**, 379-391.

Greenstone, M. H. (1984) Determinants of web spider species diversity: vegetation structural diversity vs. prey availability. *Oecologia*, **62**, 299-304.

Grill, A., Knoflach, B., Cleary, D. F. R. & Kati, V. (2005) Butterfly, spider, and plant communities in different land-use types in Sardina, Italy. *Biodiversity and Conservation*, **14**, 1281-1300.

Hatley, C. L. & Macmahon, J. A. (1980) Spider community organization: seasonal variation and the role of vegetation architecture. *Environmental Entomology*, **9**, 632-639.

Heikkinen, R. K., Luoto, M., Virkkala, R. & Rainio, K. (2004) Effects of habitat cover, landscape structure and spatial variables on the abundance of birds in an agricultural-forest mosaic. *Journal of Applied Ecology*, **41**, 824-835.

Hortal, J. (2004) *Selección y diseño de áreas prioritarias de conservación de la biodiversidad mediante sinecología. Inventario y modelización predictiva de la distribución de los escarabeidos coprófagos (Coleoptera, Scarabaeoidea) de Madrid*. PhD dissertation, Universidad Autónoma de Madrid, Madrid.

Hortal, J., Lobo, J. M. & Martín-Piera, F. M. (2001) Forecasting insect species richness scores in poorly surveyed territories: the case of the Portuguese dung beetles (Col. Scarabaeidae). *Biodiversity and Conservation*, **10**, 1343-1367.

Hortal, J., García-Pereira, P. & García-Barros, E. (2004) Butterfly species richness in mainland Portugal: Predictive models of geographic distribution patterns. *Ecography*, **27**, 68-82.

ITGE. (1988) *Atlas Geocientífico y del Medio Natural de la Comunidad de Madrid. Serie: Medio Ambiente*. Instituto Tecnológico GeoMinero de España, Madrid.

Jerardino, M., Urones, C. & Fernández, J. L. (1991) Datos ecológicos de las arañas epigeas en dos bosques de la región mediterránea. *Orsis*, **6**, 141-157.

Jiménez-Valverde, A. & Hortal, J. (2003) Las curvas de acumulación de especies y la necesidad de evaluar la calidad de los inventarios biológicos. *Revista Ibérica de Aracnología*, **8**, 151-161.

Jiménez-Valverde, A. & Lobo, J. M. (2004) Un método sencillo para seleccionar puntos de muestreo con el objeto de inventariar taxones hiperdiversos: el caso práctico de las familias *Araneidae* y *Thomisidae* (*Araneae*) en la Comunidad de Madrid, España. *Ecología*, **18**, 297-308.

Jiménez-Valverde, A. & Lobo, J. M. (2005) Determining a combined sampling procedure for a reliable estimation of Araneidae and Thomisidae assemblages (Arachnida: Araneae). *Journal of Arachnology*, **33**, 33-42.

Jiménez-Valverde, A. & Lobo, J. M. (2006) Establishing reliable spider (Araneae, Araneidae & Thomisidae) assemblage sampling protocols: estimation of species richness, seasonal coverage and contribution of juveline data to species richness and composition. *Acta Oecologica*, in press.

Jiménez-Valverde, A., Jiménez Mendoza, S., Martín Cano, J & Munguira, M. L. (2006) Comparing relative model fit of species accumulation functions to local Papilionoidea and Hesperioidea butterfly inventories of Mediterranean habitats. *Biodiversity and Conservation*, **15**, 177-190.

Lassau, S. A. & Hochuli, D. F. (2004) Effects of habitat complexity on ant assemblages. *Ecography*, **27**, 157-164.

Lassau, S. A. & Hochuli, D. F. (2005) Wasp community responses to habitat complexity in Sydney sandstone forests. *Austral Ecology*, **30**, 179-187.

Lassau, S. A., Cassis, G., Flemons, P. K. J., Wilkie, L. & Hochuli, D. F. (2005b) Using high-resolution multi-spectral imagery to estimate habitat complexity in open-canopy forests: can we predict ant community patterns? *Ecography*, **28**, 495-504.

Lassau, S. A., Hochuli, D. F., Cassis, G. & Reid, C. A. M. (2005a) Effects of habitat complexity on forest beetle diversity: do functional groups respond consistently? *Diversity and Distributions*, **11**, 73-82.

Legendre, P. & Legendre, L. (1998) *Numerical Ecology*. Elsevier, Amsterdam.

León-Cortés, J. L., Soberón-Mainero, J. & Llorente-Bousquets, J. (1998) Assessing completeness of Mexican sphinx moth inventories through species accumulation functions. *Diversity and Distributions*, **4**, 37-44.

Lobo, J. M., Jay-Robert, P. & Lumaret, J.-P. (2004) Modelling the species richness distribution for French Aphodiidae (Coleoptera, Scarabaeoidea). *Ecography*, **27**, 145-156.

Mac Nally, R. (2000) Regression and model-building in conservation biology, biogeography and ecology: the distinction between - and reconciliation of - "predictive" and "explanatory" models. *Biodiversity and Conservation*, **9**, 655-671.

Marc, P, Canard, A. & Ysnel, F. (1999) Spiders (Araneae) useful for pest limitation and bioindication. *Agriculture, Ecosystems and Environment*, **74**, 229-273.

McCoy, E. D. & Bell, S. S. (1991) Habitat structure: the evolution and diversification of a complex topic. In *Habitat structure: The Physical Arrangement of Objects in Space*, eds. S. S. Bell, E. D. McCoy & H. R. Mushinsky, pp. 3-27. Chapman and Hall, London.

New, T. R. (1999) Untangling the web: spiders and the challenges of invertebrate conservation. *Journal of Insect Conservation*, **3**, 251-256.

Newsome, A. E. & Catling, P. C. (1979) Habitat preferences of mammals inhabiting heathlands of warm temperate coastal, montane and alpine regions of southeastern Australia. In *Ecosystems of the world, Vol. 9A, Heathlands and related shrublands of the world*, ed. R. L. Specht, pp. 301-316. Elsevier, Amsterdam.

Nicholls, A. O. (1989) How to make biological surveys go further with generalised linear models. *Biological Conservation*, **50**, 51-71.

Nyffeler, M. (2000) Ecological impact of spiders predation: a critical assessment of Bristowe´s and Turnbull´s estimates. *Bulletin of the British Arachnological Society*, **11(9)**, 367-373.

Olden, J. D. & Jackson, D. A. (2000) Torturing data for the sake of generality: How valid are our regression models? *Écoscience*, **7(4)**, 501-510.

Parra, J. L., Graham, C. C. & Freile, J. F. (2004) Evaluating alternative data sets for ecological niche models of birds in the Andes. *Ecography*, **27**, 350-360.

Pascual, M. A. & Iribarne, O. O. (1993) How good are empirical predictions of natural mortality? *Fisheries Research*, **16**, 17-24.

Peterson, A. T. & Slade, N. A. (1998) Extrapolating inventory results into biodiversity estimates and the importance of stopping rules. *Diversity and Distributions*, **4**, 95-105.

Pettorelli, N., Vik, J. O., Mysterud, A., Gaillard, J.-M., Tucker, C. J. & Stenseth, N. C. (2005) Using the satellite-derived NDVI to assess ecological responses to environmental change. *Trends in Ecology and Evolution*, **20**, 503-510.

Platnick, N. I. (2005) *The World Spider Catalog v 5.5.* American Museum of Natural History, in http://research.amnh.org/entomology/spiders/catalog/INTRO1.html

Rivas-Martínez, S. (1987) *Memoria del mapa de series de vegetación de España 1:400.000*. ICONA, Madrid.

Robinson, J. V. (1981) The effect of architectural variation in habitat on a spider community: an experimental field study. *Ecology*, **62**, 73-80.

Roura-Pascual, N., Suarez, A. V., Gómez, C., Pons, P., Touyama, Y., Wild, A. L. & Peterson, A. T. (2004) Geographical potential of Argentine ants (*Linepithema humile* Mayr) in the face of global climate change. *Proceedings of the Royal Society of London B*, **271**, 2527-2535.

Ruggiero, A. & Kitzberger, T. (2004) Environmental correlates of mammal species richness in South America: effects of spatial structure, taxonomy and geographic range. *Ecography*, **27**, 401-416.

Rushton, S. P. & Eyre, M. D. (1992) Grassland spider habitats in north-east England. *Journal of Biogeography*, **19**, 99-108.

Rushton, S. P., Topping, C. J. & Eyre, M. D. (1987) The habitat preferences of grassland spiders as identified using detrended correspondence analysis (DECORANA). *Bulletin of the British Arachnological Society*, **7**, 165-170.

Rypstra, A. L. (1986) Web spiders in temperate and tropical forests: relative abundance and environmental correlates. *American Midland Naturalist*, **115(1)**, 42-51.

Rypstra, A. L., Carter, P. E., Balfour, R. A. & Marshall, S. D. (1999) Architectural features of agricultural habitats and their impact on the spider inhabitants. *Journal of Arachnology*, **27**, 371-377.

Samu, F. & Lövei, G. L. (1995) Species richness of a spider community (Araneae): extrapolation from simulated increasing sampling effort. *European Journal of Entomology*, **92**, 633-638.

Soberón, J. & Llorente, B. J. (1993) The use of species accumulation functions for the prediction of species richness. *Conservation Biology*, **7**, 480-488.

Sørensen, L. L. (2004) Composition and diversity of the spider fauna in the canopy of a montane forest in Tanzania. *Biodiversity and Conservation*, **13**, 437-452.

Sørensen, L. L., Coddington, J. A. & Scharff, N. (2002) Inventorying and estimating subcanopy spider diversity using semiquantitative sampling methods in an Afromontane forest. *Environmental Entomology*, **31**, 319-330.

StatSoft (2001) *STATISTICA (data analysis software system and user´s manual). Version 6.* StatSoft, Inc., Tulsa, OK.

Suárez-Seoane, S., Osborme, P. E. & Alonso, J. C. (2002) Large-scale habitat selection by agricultural steppe birds in Spain: identifying species-habitat responses using generalizad additive models. *Journal of Applied Ecology*, **39**, 755-771.

Tews, J., Brose, U., Grimm, V., Tielbörger, K. Wichmann, M. C., Schwager, M. & Jeltsch, F. (2004) Animal species diversity driven by habitat heterogeneity/diversity: the importance of keystone structures. *Journal of Biogeography*, **31**, 79-92.

Toti, D. S., Coyle, F. A. & Miller, J. A. (2000) A structured inventory of Appalachian grass bald and heath bald spider assemblages and a test of species richness estimator performance. *Journal of Arachnology*, **28**, 329-345.

Turnbull, A. L. (1973) Ecology of true spiders (Araneomorphae). *Annual Review of Entomology*, **18**, 305-348.

Uetz, G. W. (1991) Habitat structure and spider foraging. In *Habitat structure: The Physical Arrangement of Objects in Space,* eds. S. S. Bell, E. D. McCoy & H. R. Mushinsky, pp. 325-348. Chapman and Hall, London.

Urones, C. & Puerto, A. (1988) Ecological study of the Clubionoidea and Thomisoidea (Araneae) in the Spanish Central System. *Revue Arachnologique*, **8(1)**, 1-32.

Wise, D.H. (1993) *Spiders in ecological webs*. Cambridge University Press, New York.

**Segunda Parte: Distribución potencial del endemismo ibérico *Macrothele calpeiana* (Walckenaer, 1805) (Araneae, Hexathelidae)**

# CRITERIOS PARA SELECCIONAR EL PUNTO DE CORTE CON EL FIN DE CONVERTIR MAPAS CONTINUOS DE PROBABILIDAD DE PRESENCIA A MAPAS BOOLEANOS DE PRESENCIA/AUSENCIA

**RESUMEN.** Para convertir un mapa continuo de probabilidades de presencia a uno booleano de presencia/ausencia es necesario fijar un valor de probabilidad por encima del cual considerar a la especie como presente. Debido al sesgo presente en las probabilidades logísticas por el efecto de la prevalencia, un punto de corte fijo, como 0.5, no suele corresponder con el punto de corte por encima del cual es más probable encontrar a la especie. En este trabajo se comparan, para diversos tamaños de muestra y prevalencias, cuatro criterios para seleccionar un punto de corte, modelizando una especie virtual con el fin de evitar las fuentes de error que inevitablemente existen en los datos reales. En general, los criterios que minimizan la diferencia o maximizan la suma entre la sensitividad y la especificidad son los que producen las mejores predicciones, especialmente el primero y en el caso de trabajar con especies raras, aquellas de mayor interés conservacionista. Los criterios del 0.5 y de maximizar el estadístico Kappa son los peores en prácticamente todas las situaciones. Cualquiera que sea el criterio usado para fijar el punto de corte, tanto el valor del punto como las razones que han determinado su elección debe ser siempre especificado.

**Palabras clave:** matriz de confusión, modelos predictivos de distribución, estadístico Kappa, regresión logística, punto de corte

**Este capítulo ha sido enviado a publicar como:**

JIMÉNEZ-VALVERDE, A. & LOBO, J. M. Threshold criteria for conversion of probability of species presence to either-or presence-absence. *Acta Oecologica.*

# THRESHOLD CRITERIA FOR CONVERSION OF PROBABILITY OF SPECIES PRESENCE TO EITHER-OR PRESENCE/ABSENCE

**ABSTRACT.** The continuous map of probability of presence produced by logistic regressions is converted into an either-or presence/absence map, so a threshold probability indicative of species presence must be fixed. Because of the bias in probability outputs due to frequency of presences (prevalence), a fixed threshold value, such as 0.5, does not usually correspond to the threshold above which the species is more likely to be present. In this paper four decision threshold criteria are compared for a wide range of sample sizes and prevalences, modeling a virtual species in order to avoid the omnipresent error sources that the use of real species data implies. In general, sensitivity-specificity difference minimizer and sensitivity-specificity sum maximizer criteria produced the most accurate predictions, especially the former and in the case of rare species, which are the ones with the greatest conservation interest. The widely-used 0.5 fixed threshold and Kappa-maximizer criteria are the worst ones in almost all situations. Nevertheless, whatever the criteria used, the threshold value chosen and the research goals that determined its choice should be stated in works that purport to further biodiversity conservation.

**Keywords:** confusion matrix, habitat-suitability models, Kappa statistic, logistic regression, threshold

## INTRODUCTION

Species distribution is increasingly being modelled in ecology and conservation research. Prediction of species geographic distribution, based on known occurrences, is now possible thanks to both Geographic Information Systems (GIS) and statistical quantification of species-environment relationships (Guisan & Zimmermann, 2000; Lehmann *et al*., 2002; Rushton *et al*., 2004). Habitat model predictions help to delve into questions of biogeography and evolution (Peterson *et al*., 1999; Anderson *et al*., 2002a, b; Peterson & Holt, 2003), to search for biological indicators (Bonn & Schröder, 2001), to study the effect of climate warming on species distribution (Teixeira & Arntzen, 2002), and to develop management decisions and conservation strategies (Godown & Peterson, 2000; Schadt *et al*., 2002; Barbosa *et al*., 2003; Meggs *et al*., 2004; Russell *et al*., 2004; Chefaoui *et al*., 2005).

Prediction methods currently available to scientists can be divided, roughly, into those that use only presence data (profile techniques, e.g. environmental envelopes) and those that also incorporate absence data (group discrimination techniques, e.g. generalized regression, see Guisan & Zimmermann, 2000 and Scott *et al*., 2002). Profile techniques tend, in general, to overestimate distributions due to the lack of absence data, which would otherwise correct predictions if included (Ferrier & Watson, 1997; Zaniewski *et al*., 2002; Engler *et al*., 2004). Reliable absence data should be treated with group discrimination techniques, capable of accounting for more useful relationships between species and environmental and historical factors (Hirzel *et al*., 2001; Brotons *et al*., 2004; Segurado & Araújo, 2004). Logistic regression (LR), belonging to the generalized linear model family (GLM), in which the estimated probability of occurrence of an event is predicted as a function of one or more

independent variables, is widely used in ecology studies (Guisan *et al*., 2002; Lehmann *et al*., 2002; Reineking & Schröder, 2003). As an extension of regression methods, the LR technique, although not without its problems (Huston, 2002), is generally robust and reliable, easily executed on the majority of available statistical software and easily implemented with GIS.

The continuous map of probability of presence produced by LR techniques is converted into an either-or presence/absence map for two main reasons: first, to evaluate model prediction reliability, involving comparison with the inherently either-or presence/absence data using a confusion matrix; second, to provide categorical maps which are useful for many practical applications. This conversion involves adopting a threshold probability indicative of species presence (Fielding & Bell, 1997) which will determine model output, as it will condition the cases assigned to each category (Fielding & Bell, 1997; Manel *et al*., 1999b; Pearce & Ferrier, 2000a). However, logistic regression probabilities are biased toward the highest number of either presences or absences, where they differ (Hosmer & Lemeshow, 1989; Cramer, 1999). Because of this bias, due to prevalence (the proportion of presence cases), the threshold value of 0.5, often adopted (e. g. Li *et al*., 1997; Manel *et al*., 1999; Fleishman *et al*., 2003; Berg *et al*., 2004; Meggs *et al*., 2004) does not actually correspond to the threshold above which the species is more likely to be present. For example, where a large number of target-species absence observations bias probabilities toward zero, a cut-off of 0.5 will lead to absence predictions for sites with known presences (high omission error rate), reduce sensitivity (true predicted presences) and increase specificity (true predicted absences). Lowering the threshold from 0.5 would increase sensitivity, at the expense of decreased specificity. How to choose the best threshold for binary data with a dissimilar number of presences and absences? Although prevalence

seems to influence cut-off selection most strongly, the number of observations can also influence model performance (Pearce & Ferrier, 2000b); thresholds selected for small sample sizes can produce misleading presence/absence maps. Should sample size affect threshold choice?

The choice of threshold criteria can depend on the role of commission (false positive) and omission (false negative) errors (Fielding & Bell, 1997; Fielding, 2002; Pearson *et al*., 2004). However, models usually are designed to discriminate as reliably as possible between presence and absence sites. One study (Manel *et al*., 2001) of a large set of species concludes that results from a threshold which maximized the sensitivity-specificity sum (following Zweig & Campbell, 1993) were superior to results from a threshold of 0.5, after comparison of the two. Liu *et al*. (2005) compared 12 threshold decision criteria using data of two plant species in Europe modelled with neural networks and, although the technique used differs considerably from regression approaches, their conclusions are interesting: fixed thresholds and those based on the Kappa statistic work worse that those accounting, directly or indirectly, for prevalence.

The present study compares model outputs obtained from varying sample size and prevalence data, modeled with LR, and four widely-accepted threshold criteria: 0.5 cut-off; Kappa maximization (e. g. Guisan & Hofer, 2003; Engler *et al*., 2004; Segurado & Araújo, 2004); sensitivity-specificity difference minimizer (e. g. Bonn & Schröder, 2001; Barbosa *et al*., 2003); and sensitivity-specificity sum maximizer (Manel *et al*., 2001). The general aim of this paper is to find the optimum threshold criteria for a wide range of model specifications. In order to achieve this objective real data are not used to avoid the frequently error sources that their use implies; instead, a distribution of a virtual species was postulated.

## METHODS

**The virtual species. —** As has been recommended (Allredge & Ratti, 1986), predictions derived from four threshold criteria were compared using a postulated species distribution with known environmental influence. This procedure has been employed by other resarchers (Hirzel *et al*., 2001; Reese *et al*., 2005) to avoid complications from natural variation. Specifically, we have used this approach in order to:

i) Be sure that the modeling technique (LR) can correctly predict species distribution while avoiding the bias due to contingent unaccounted-for or unknown factors. Model distribution predictions were based on explanatory variables also used as environmental variables to build the distribution of the species.

ii) Eliminate the random noise always present in biological data, which can be modelled due to overfitting.

iii) Be completely confident about models accuracy. Many authors suggest that model performance should be evaluated using data independent from that used to generate predictions, but genuine distribution data can not usually be used to test a model if we want to use all the available information to train it. Original data from a postulated, virtual species can be used to test prediction power of models.

The virtual species distribution was mapped at a spatial resolution of 0.04º degrees for the European region (-13º to 35º longitude, and 34º to 72º latitude). The total area of the region studied measured 6576424 $km^2$ (510514 squares). For this region,

four environmental variables were extracted from WORLDCLIM interpolated map database (version 1.3; see http://biogeo.berkeley.edu/worldclim/worldclim.htm): total annual precipitation, summer precipitation, mean maximum temperature and mean minimum temperature. Box-Cox normalized environmental variables were standardized to 0 mean and 1 standard deviation to eliminate measurement-scale effects. Principal Component Analysis (PCA; see Legendre & Legendre, 1998) was performed to obtain two reduced non-correlated environmental variables able to explain 92.6% of the environmental variation across the European region: one positively correlated with temperature variables, and a second correlated with precipitation variables. The mean scores of these two environmental factors were calculated. The environment range inhabited by the species was set to the mean ± SD of each factor. All cells falling within these intervals for both factors were selected as the true distribution range of the virtual species in Europe (presences; *n*=91144), while the remaining cells were considered as true absences (*n*=419296). All geographic analysis was done with IDRISI Kilimanjaro software (Clark Labs, 2003). The geographical distribution of this "central" European virtual species (Fig. 1) is completely conditioned by well-known environmental factors.

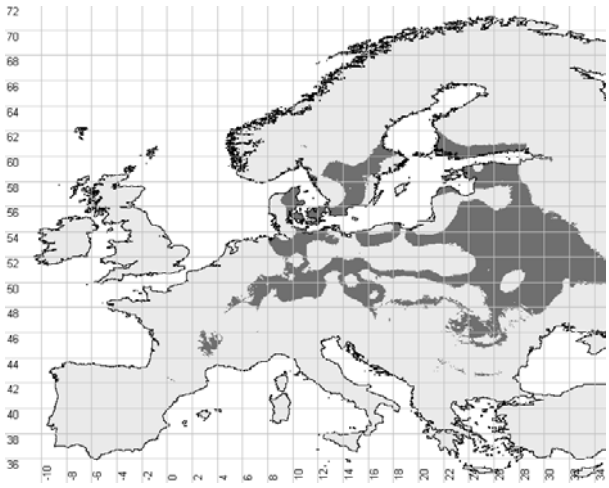


Figure 1.- Distribution of the "central European" virtual species. Its geographic range is entirely determined by *a priori* selected environmental variables.

**The modeling process. —** Nine categories of presence numbers (*n*=91, 456, 911, 4557, 9114, 22786, 45572, 68358 and 91144) were randomly selected from the distribution, correspond with successive increased percentages of presences (0.1%, 0.5%, 1%, 5%, 10%, 25%, 50%, 75% and 100%). Absences were also randomly selected in the same number categories as presences. Thus 81 models were designed, which vary both in the number of observations (from 182 to 182288) and in the prevalence or proportion of presences (from 0.001 to 0.999). All models were designed using logistic regression analysis (StatSoft, 2001) with a binomial error distribution. The cubic functions of the two environmental factors, together with their interaction, were used as explanatory variables; models were selected by a backward stepwise procedure.

Statistics to compare the accuracy of two raster maps are derived from a cross-tabulated matrix of the number of observed presence and absence cases against the predicted presences and absences (confusion matrix). Commission errors (model predictions of species presence where not actually observed, i.e. the false positive fraction) and omission errors (model predictions of species absence where actually observed; i.e. the false negative fraction) are determined by the number of cases correctly and incorrectly assigned to presences and absences. Specificity is calculated as the ratio of correctly predicted absences to the total number of absences, and sensitivity as the ratio of correctly predicted presences to their total number. The models were then projected onto the whole European territory and their probability scores converted into a binary variable (presence/absence) by applying the threshold criteria explained in the next section, based on the formerly described confusion matrix. Predicted and real virtual maps were compared by calculating the sensitivity and specificity as well as the

frequently-used Kappa statistic (Monserud & Leemans, 1992; Fielding & Bell, 1997; Pontius, 2000) that takes into consideration both commission and omission errors.

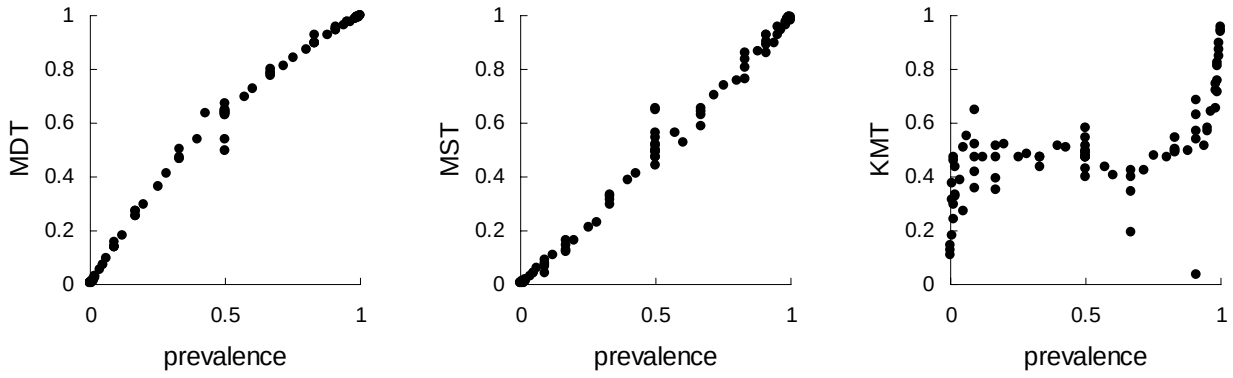
**Threshold criteria. —** Model extrapolations were converted into presence/absence maps by selecting threshold probabilities above which presence was established, according to the following criteria:

- 0.5T criteria: a value of 0.5 was the threshold above which presence was assigned.
- KMT criteria (Kappa-maximized threshold): Kappa scores were calculated for 100 threshold values (in 0.01 increments) and the one which provides maximum Kappa became the accepted threshold.
- MDT criteria (minimized difference threshold): difference between sensitivity and specificity was calculated for the same 100 threshold values and the one which minimized that difference was selected.
- MST criteria (maximized sum threshold): sum of sensitivity and specificity was calculated for the same 100 threshold values and the one which maximized that sum was selected.

## RESULTS

While sample size is uncorrelated with the thresholds selected by MDT, MST and KMT criteria (Spearman rank correlation coefficients, $r$=-0.03, -0.02 and 0.04, respectively), prevalence is significantly and positively correlated with them. Both MDT and MST thresholds are linearly related with prevalence (Fig. 2), so frequency of presence data alone could be used to select the most appropriate cut-off. However, KMT thresholds increase rapidly with either low or high prevalence values, remaining

relatively constant (around 0.5) in a wide range of prevalence values. The thresholds from the KMT, MDT and MST criteria are also correlated with each other, being MDT and MST thresholds highly positively correlated (Fig. 2).




| | Prevalence | MDT | KMT |
|---|---|---|---|
| MDT | 0.99 | - | - |
| KMT | 0.73 | 0.73 | - |
| MST | 0.99 | 0.99 | 0.74 |

Figure 2.- Relationship between prevalence of occurrence and threshold scores for the three non-fixed cut-off criteria. In the table, Spearman correlation scores are shown between prevalence and threshold values and between threshold scores of the three criteria.

The four threshold criteria produced significantly different mean values of sensitivity, specificity and Kappa (Table 1). Mean sensitivity and Kappa values were significantly higher for MST and MDT, while they were significantly lower for 0.5T and KMT. Mean specificity values were significantly higher only for MDT (Table 1).

Table 1.- Mean values of the three accuracy measures (Kappa statistic, sensitivity and specificity) ± SE for the four cut-off threshold criteria, and ANOVA results (*** < 0.001, * < 0.005). Pairwise significant differences were determined using a Tukey test (HDS; $p < 0.05$) and are shown with letters.

| | **Kappa** | **Sensitivity** | **Specificity** |
|---|---|---|---|
| $F_{(3, 320)}$ | 29.09 *** | 10.97*** | 4.48 * |
| [a] MST | 0.734 ± 0.003 [b, d] | 0.956 ± 0.002 [b, d] | 0.898 ± 0.002 [c] |
| [b] 0.5T | 0.599 ± 0.024 [a, c] | 0.799 ± 0.035 [a, c] | 0.894 ± 0.009 [c] |
| [c] MDT | 0.766 ± 0.002 [b, d] | 0.926 ± 0.002 [b, d] | 0.923 ± 0.001 [a, b, d] |
| [d] KMT | 0.644 ± 0.016 [a, c] | 0.833 ± 0.028 [a, c] | 0.898 ± 0.008 [c] |

Kappa and specificity values obtained with MST and MDT (good-performance cut-offs) are significantly correlated, as are those obtained with 0.5T and KMT thresholds (poor-performance thresholds); the latter producing the highest correlation scores (Table 2). Sensitivity values are also positively correlated for the pairs MST/KMT and MST/0.5T, although again 0.5T and KMT producing the highest correlation score. Sensitivity scores from 0.5T and KMT seemingly linearly related (Fig. 3), were extremely variable in comparison with those from MST and MDT, which, while correlated, were always higher than 0.8. The pattern for the Kappa statistic is quite similar, while specificity values were high for the four criteria.

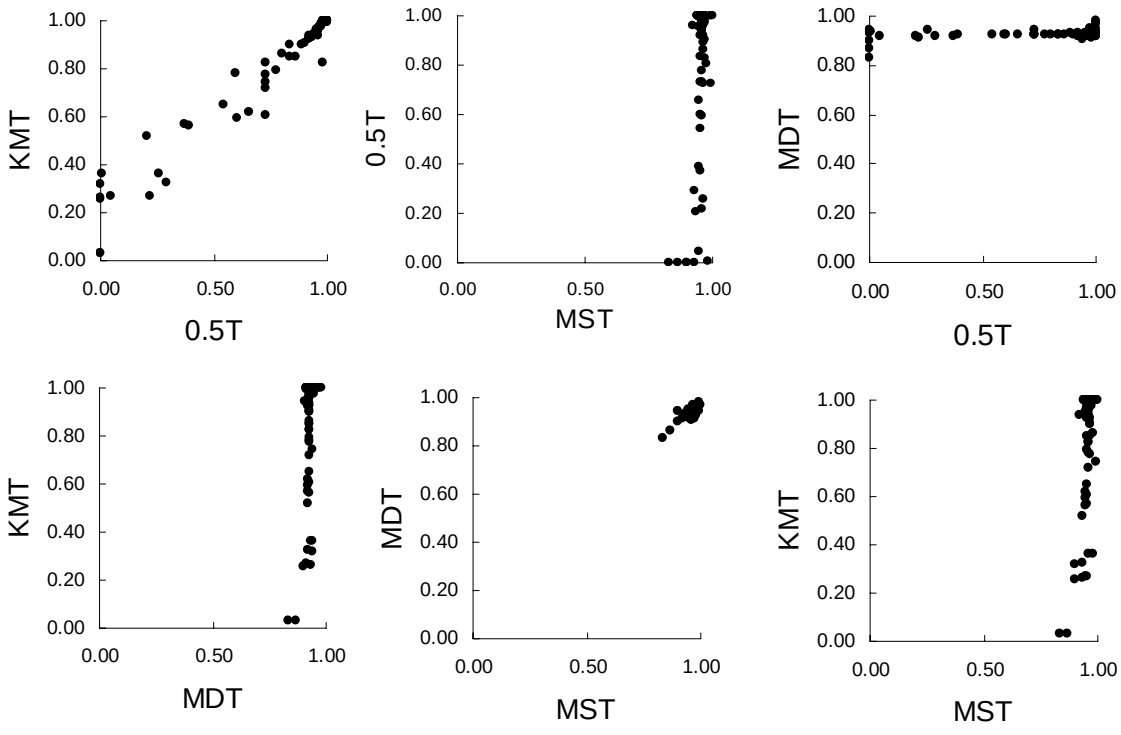



Figure 3.- Sensitivity scores obtained using different cut-off criteria. While those get with 0.5T and KMT criteria are highly variable, sensitivity values obtained with MDT and MST criteria show always quite high values.

Table 2.- Spearman correlation scores between cut-off criteria for the three accuracy measures, applying Bonferroni correction for multiple comparison test. Significant pairwise are marked in bold.

| | **Kappa** | **Sensitivity** | **Specificity** |
|---|---|---|---|
| MST/0.5T | 0.06 | **0.42** | 0.16 |
| MST/MDT | **0.52** | **0.40** | **0.35** |
| MST/KMT | 0.02 | **0.46** | 0.18 |
| 0.5T/MDT | 0.33 | 0.24 | -0.04 |
| 0.5T/KMT | **0.93** | **0.98** | **0.98** |
| MDT/KMT | 0.21 | 0.27 | -0.07 |

Kappa, sensitivity and specificity score differences derived from the various threshold criteria (relative performance) cannot be explained by variation in sample size, but can be explained by variation in prevalence (Table 3), which explains (with the exception of MDT-MST) 41% to 98% of their deviance. Regardless of the grouping of criteria into good-performers (MDT and MST) and poor-performers (0.5T and KMT), the relative performance of all four criteria varies with prevalence. The reliability of MDT- or MST-criteria designed presence-absence maps proved to be independent of the frequency of presence points. However, poor-performance thresholds predicted presences relatively reliably in cases of high prevalence scores, and absences relatively reliably at low prevalence scores (Fig. 4). However, in such cases 0.5T and KMT criteria superiority over MDT and MST is negligible. It is interesting to highlight the MDT- and MST-criteria significant superiority over the other two in predicting presences when the prevalence is low (Fig. 4B); no such pattern was observed in the prediction of absences.

Table 3.- The relative difference in accuracy using different cut-off criteria was modeled using Generalized Linear Models (GLMs) with a log-link function and a normal distribution. The independent variables (sample size and prevalence of presence data) were included in the model considering their cubic, quadratic and linear functions, and the adequacy of the models was tested by means of the change in explained deviance (% Expl. Dev.) from a null model in which the difference in relative performance is modelled with no explanatory variables.

| | **Sample-size** | **Prevalence** | |
|---|---|---|---|
| | % Exp. Dev. | % Exp. Dev. | Function |
| **Kappa** | | | |
| MDT-0.5T | 0.05 | 95.94 | cubic |
| MDT-KMT | 0.24 | 82.21 | cubic |
| MST-0.5T | 0.03 | 95.97 | cubic |
| MST-KMT | 0.26 | 85.54 | cubic |
| MDT-MST | 0.46 | 7.30 | cubic |
| **Sensitivity** | | | |
| MDT-0.5T | 0.08 | 96.38 | cubic |
| MDT-KMT | 0.01 | 93.31 | cubic |
| MST-0.5T | 0.09 | 97.75 | cubic |
| MST-KMT | 0.01 | 95.19 | cubic |
| MDT-MST | <0.01 | 12.94 | quadratic |
| **Specificity** | | | |
| MDT-0.5T | <0.01 | 74.46 | linear |
| MDT-KMT | <0.01 | 69.44 | quadratic |
| MST-0.5T | <0.01 | 54.93 | linear |
| MST-KMT | 0.02 | 40.68 | linear |
| MDT-MST | 0.38 | 8.75 | cubic |

## DISCUSSION

Prediction reliability from models is particularly sensitive to threshold criteria applied in model derivation. Results herein, derived from a wide range of conditions, providing some guidance to the choice of threshold criteria, recommend above all that threshold criteria should be dependent on prevalence.

Mean LR probability magnitudes, biased by prevalence, tend toward zero for rare species (narrow geographic range, i.e. low prevalence scores) and toward one for common species (widespread, i. e. high prevalence scores). Thus, as shown in the

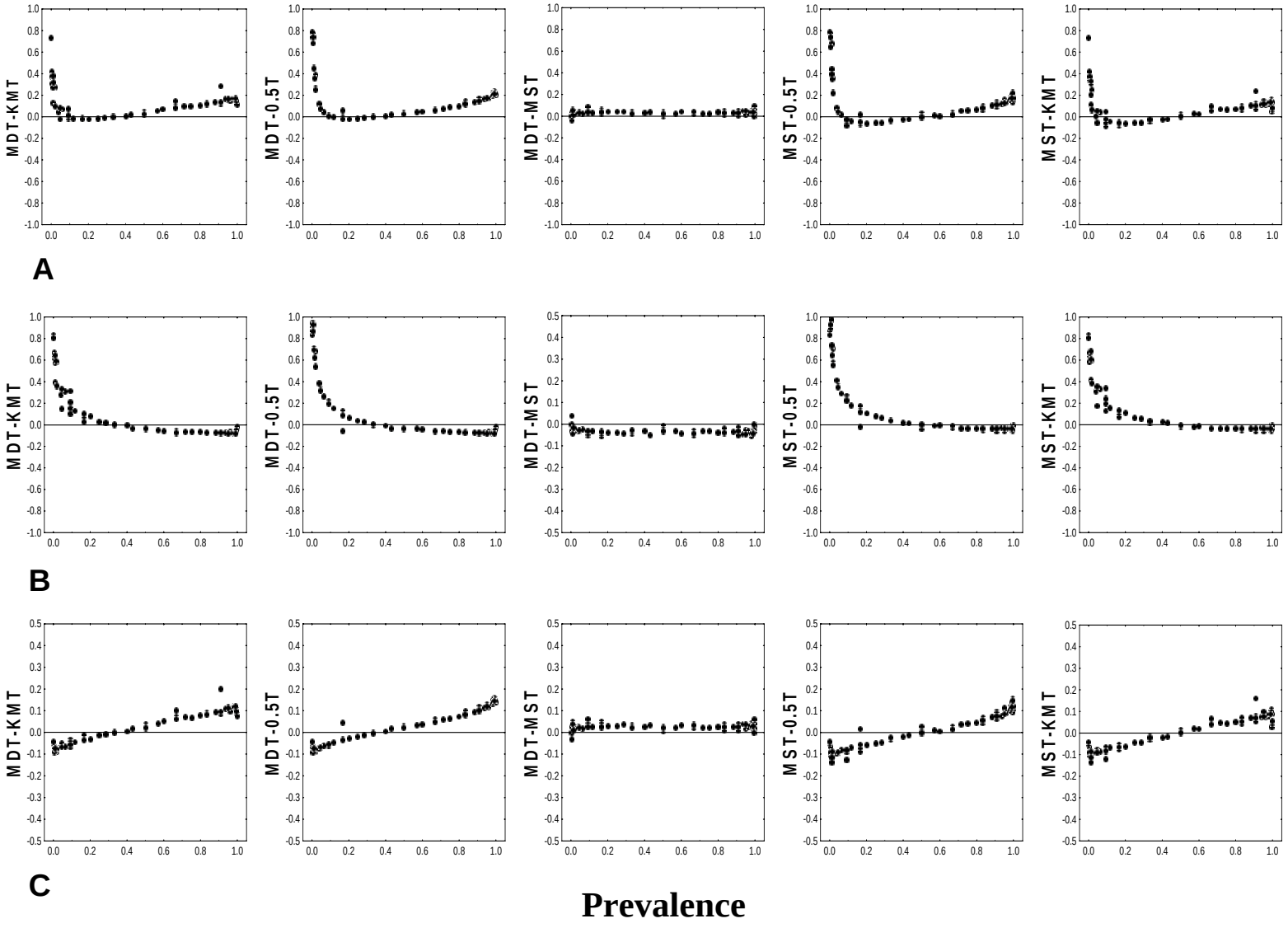



Figure 4.- Relative different performance between good-performance (MDT and MST) and bad-performance (KMT and 0.5T) critera along the prevalence of occurrence gradient (A, Kappa statistic; B, sensitivity; C, specificity).

present work, a threshold fixed a priori yields a binary model that is not biologically meaningful. Of the KMT, MDT and MST criteria tested, the ones which maximize the sensitivity-specificity sum or minimize their difference (MST and MDT, respectively) are the most linearly related with prevalence, a desirable property since the meaningfulness of model probabilities depends on the maximum value obtained (Pontius & Batchu, 2003). The strong correlation between the threshold values from these two criteria and prevalence obtained by us supports the recent proposal of using prevalence values themselves as threshold decision criteria (Liu *et al*., 2005), as previously recognized by statisticians (Cramer, 1999).

These two prevalence-dependent thresholds, strongly correlated, always score high in accuracy. KMT criteria produce quite variable accuracy scores, highly correlated with 0.5T scores, a consequence of the stability of the threshold value around 0.5 in a wide range of prevalence conditions.

Although MDT- and MST-criteria model predictions are, in general, significantly more accurate, KMT and 0.5T can be used in some circumstances: i) when accurateness in predicting presences is the objective and the prevalence is high, and ii) when we want to predict absences but the prevalence is low. If the Kappa statistic is used to measure model accuracy, then KMT and 0.5T criteria outperform in a prevalence interval of 0.1-0.5, approximately. Nevertheless, their performance differs only negligibly with respect to MDT and MST.

The measurement and meaningfulness of accuracy estimations depends on the purpose of the research, leading to varying concerns about accuracy. For example, a cut-off threshold optimizing species absences may lead to a suboptimal classification when omission errors are undesirable. While it is frequently assumed that commission and omission errors are equal costwise, in conservation it is probably more costly to classify

a recognized presence site as absence than vice versa (Fielding, 2002). Omission errors should therefore be avoided and sensitivity favoured; in this situation, MDT and MST are the threshold criteria that should be employed. On the contrary, if commission errors are considered more costly, MDT is the only criteria which produce higher specificity values. Hence, we recommend the MDT threshold criteria as the one of more general use (see also Liu *et al*., 2005 and Jiménez-Valverde & Lobo, 2006).

It is important to highlight MDT- and MST-criteria superiority over the other two in predicting presences when the prevalence of presences is low. Low prevalence scores usually characterize rare species, which are of special relevance in conservation. Moreover, as rarity is a widespread pattern in nature (Gaston, 2003), there will usually be more absence that presence points. Other study conclusions question the reliability of predictions when the MST criterion is used. Manel *et al*. (2001) extrapolated their models for aquatic invertebrates to Himalayan regions different from those where the models were trained and observed that potential distributions of rare species were overestimated. However, we think such overestimation is probably not due to the threshold criteria chosen, but to the loss of accuracy of models extrapolated to areas different from those used in model design (Fielding & Haworth, 1995; Marsden & Fielding, 1999). Thus, we recommend MDT or MST criteria as especially optimal procedures for rare species model predictions

When the cost of omission relative to commission errors can be defined *a priori*, then threshold criteria choice can be based on costs of false presences relative to false absences (Fielding & Bell, 1997; Fielding, 2002). In such a case, the relative performance of MDT and MST criteria is a line of future research which promises interesting results.

On occasion, some authors have failed to point out the threshold used (e.g. Teixeira & Arntzen, 2002), a practice which should be avoided. The criteria employed for deciding the threshold, whatever it is, as well as the threshold value itself, should be specified so that readers can reach their own conclusions.

## ACKNOWLEDGEMENTS

This paper has been supported by a Fundación BBVA project (Diseño de una red de reservas para la protección de la Biodiversidad en América del sur austral utilizando modelos predictivos de distribución con taxones hiperdiversos) and a MEC Project (CGL2004-04309), as well as by a Museo Nacional de Ciencias Naturales/C.S.I.C./Comunidad de Madrid PhD grant.

## LITERATURE CITED

Allredge, J. R. & Ratti, J. T. (1986) Comparison of some statistical techniques for analysis of resource selection. *Journal of Wildlife Management*, **50,** 157-165.

Anderson, R. P., Gómez-Laverde, M. & Peterson, A. T. (2002a) Geographical distributions of spiny pocket mice in South America: insights from predictive models. *Global Ecology and Biogeography*, **11,** 131-141.

Anderson, R. P., Peterson, A. T. & Gómez-Laverde, M. (2002b) Using niche-based GIS modeling to test geographic predictions of competitive exclusión and competitive release in South American pocket mice. *Oikos*, **98**, 3-16.

Barbosa, A. M., Real, R., Olivero, J. & Vargas, J. M. (2003) Otter (*Lutra lutra*) distribution modeling at two resolution scales suited to conservation planning in the Iberian Peninsula. *Biological Conservation*, **114**, 377-387.

Bonn, A. & Schröder, B. (2001) Habitat models and their transfer for single and multi species groups: a case study of carabids an an alluvial forest. *Ecography*, **24**, 483-496.

Brotons, L., Thuiller, W., Araújo, M. B. & Hirzel, A. H. (2004) Presence-absence versus presence-only modelling methods for predicting bird habitat suitability. *Ecography*, **27**, 437-448.

Chefaoui, R. M., Hortal, J. & Lobo, J. M. (2005) Potential distribution modelling, niche characterization and conservation status assessment using GIS tools: a case study of Iberian *Copris* species. *Biological Conservation*, **122**, 327-338.

Clark Labs (2003) *Idrisi Kilimanjaro. GIS software package*. Clark Labs, Worcester, MA.

Cramer, J. S. (1999) Predictive performance of binary logit model in unbalanced samples. *Journal of the Royal Statistical Society: Series D (The Statistician)*, **48**, 85-94.

Engler, R., Guisan, A & Rechsteiner, L. (2004) An improved approach for predicting the distribution of rare and endangered species from occurrence and pseudo-absence data. *Journal of Applied Ecology*, **41**, 263-274.

Ferrier, S., & Watson, G. (1997) *An evaluation of the effectiveness of environmental surrogates and modelling techniques in predicting the distribution of biological diversity.* Environment Australia, Canberra, available in http://www.deh.gov.au/biodiversity/publications/technical/surrogates/

Fielding, A. H. (2002) What are the appropiate characteristics of an accuracy measure? In *Predicting Species Occurrences. Issues of Accuracy and Scale*, eds. J. M. Scott, P. J. Heglund, J. B. Haufler, M. Morrison, M. G. Raphael, W. B. Wall & F. Samson, pp. 271-280. Island Press, Covelo, CA,

Fielding, A. H. & Bell, J. F. (1997) A review of methods for the assessment of prediction errors in conservation presence/absence models. *Environmental Conservation*, **24**, 38-49.

Fielding, A. H. & Haworth, P. F. (1995) Testing the generality of bird-habitat models. *Conservation Biology*, **9**, 1466-1481.

Fleishman, E., Mac Nally, R., Fay, J. P. & Murphy, D. D. (2001) Modeling and predicting species occurrence using broad scale environmental variables: an example with butterflies of the Great Basin. *Conservation Biology*, **15**, 1674-1685.

Gaston, K. J. (2003) *The structure and dynamics of geographic ranges*. Oxford University Press, Oxford.

Guisan, A. & Zimmermann, N. E. (2000) Predictive habitat distribution models in ecology. *Ecological Modelling*, **135**, 147-186.

Guisan, A., Edwards, T. C. & Hastie, T. (2002) Generalized linear and generalized additive models in studies of species distributions: setting the scene. *Ecological Modelling*, **157**, 89-100.

Guisan, A. & Hofer, U. (2003) Predicting reptile distributions at the mesoscale: relation to climate and topography. *Journal of Biogeography*, **30**, 1233-1243.

Godown, M. & Peterson, A. T. (2000) Preliminary distributional analysis of US endangered bird species. *Biodiversity and Conservation*, **9**, 1313-1322.

Hirzel, A. H., Helfer, V. & Metral, F. (2001) Assessing habitat-suitability models with a virtual species. *Ecological Modelling*, **145**, 111-121.

Hosmer, D. W. & Lemeshow, S. (1989) *Applied Logistic Regression*. Wiley, New York.

Huston, M.A. (2002) Introductory essay: Critical issues for improving predictions. In *Predicting Species Occurrences. Issues of Accuracy and Scale*, eds. J. M. Scott, P. J. Heglund, J. B. Haufler, M. Morrison, M. G. Raphael, W. B. Wall & F. Samson, pp. 7-21. Island Press, Covelo, CA,

Jiménez-Valverde, A. & Lobo, J. M. (2006) The ghost of unbalanced species distribution data in geographical model predictions. *Diversity and Distributions*, in press.

Legendre, P. & Legendre, L. (1998) *Numerical Ecology*. Elsevier, Amsterdam.

Lehmann, A., Overton, J. McC. & Austin, M. P. (2002) Regression models for spatial prediction: their role for biodiversity and conservation. *Biodiversity and Conservation*, **11**, 2085-2092.

Li, W., Wang, Z., Ma, Z. & Tang, H. (1997) A regression model for the spatial distribution of red-crown crane in Yancheng Biosphere Reserve, China. *Ecological Modelling*, **103**, 115-121.

Liu, C., Berry, P. M., Dawson, T. P. & Pearson, R. G. (2005) Selecting thresholds of accurrence in the prediction of species distributions. *Ecography*, **28**, 385-393.

Manel, S., Dias, J. M., Buckton, S. T. & Ormerod, S. J. (1999) Alternative methods for predicting species distributions: an illustration with Himalayan river birds. *Journal of Applied Ecology*, **36**, 734-747.

Manel, S., Dias, J. M., Buckton, S. T. & Ormerod, S. J. (1999b) Comparing discriminat analysis, neural networks and logistic regression for predicting species distributions: a case study with a Himalayan river bird. *Ecological Modelling*, **120**, 337-347.

Manel, S., Williams, H. C. & Ormerod, S. J. (2001) Evaluating presence-absence models in ecology: the need to account for prevalence. *Journal of Applied Ecology*, **38**, 921-931.

Marsden, S. & Fielding, A. (1999) Habitat associations of parrots on the Wallacean islands of Buru, Seram and Sumba. *Journal of Biogeography*, **26**, 439-446.

Meggs, J. M., Munks, S. A., Corkrey, R. & Richards, K. (2004) Development and evaluation of predictive habitat models to assist the conservation planning of a

threatened lucanid beetle, *Hoplogonus simsoni*, in north-east Tasmania. *Biological Conservation*, **118**, 501-511.

Monserud, R. A. & Leemans, R. (1992) Comparing global vegetation maps with the Kappa statistic. *Ecological Modelling*, **62**, 275-293.

Pearce, J. & Ferrier, S. (2000a) Evaluating the predictive performance of habitat models developed using logistic regression. *Ecological Modelling*, **133**, 225-245.

Pearce, J. & Ferrier, S. (2000b) An evaluation of alternative algorithms for fitting species distribution models using logistic regression. *Ecological Modelling*, **128**, 127-147.

Pearson, R. G., Dawson, T. P. & Liu, C. (2004) Modelling species distributions in Britain: a hierarchical integration of climate and land-cover data. *Ecography*, **27**, 285-298.

Peterson, A. T. & Holt, R. D. (2003) Niche differentiation in Mexican birds: using point occurrences to detect ecological innovation. *Ecology Letters*, **6**, 774-782.

Peterson, A. T., Soberón, J. & Sánchez-Cordero, V. (1999) Conservatism of ecological niches in evolutionary time. *Science*, **285**, 1265-1267.

Pontius, R. G. (2000) Quantification error versus location error in comparison of categorical maps. *Photogrammetric Engineering & Remote Sensing*, **66**, 1011-1016.

Pontius, R. G. & Batchu, K. (2003) Using the Relative Operating in prediction of location of land cover change in India. *Transactions in GIS*, **7**, 467-484.

Reese, G. C., Wilson, K. R., Hoeting, J. A. & Flather, C. H. (2005) Factors affecting species distribution predictions: a simulation modeling experiment. *Ecological Applications*, **15**, 554-564.

Reineking, B. & Schröder, B. (2003) Computer-intensive methods in the analysis of species-habitat relationships. In *GfÖ Arbeitskreis Theorie in der Ökologie: Gene, Bits und Ökosysteme*, ed. H. Reuter, B. Breckling & A. Mittwollen, pp. 100-117. P. Lang Verlag Frankfurt/M.

Rushton, S. P., Ormerod, S. J. & Kerby, G. (2004) New paradigms for modelling species distributions? *Journal of Applied Ecology*, **41**, 193-200.

Russell, K. R., Mabee, T. J. & Cole, M. B. (2004) Distribution and habitat of Columbia Torrent Salamanders at multiple spatial scales in managed forests of Northwestern Oregon. *Journal of Wildlife Management*, **68**, 403-415.

Schadt, S., Revilla, E., Wiegand, T., Knauer, F., Kaczensky, P., Breitenmoser, U., Bufka, L., Červený, J., Koubek, P., Huber, T., Staniša, C. & Trepl, L. (2002) Assessing the suitability of central European landscapes for the reintroduction of Eurasian lynx. *Journal of Applied Ecology*, **39**, 189-203.

Scott, J. M., Heglund, P. J., Haufler, J. B., Morrison, M., Raphael, M. G., Wall, W. B. & Samson, F. (eds) (2002) *Predicting Species Occurrences. Issues of Accuracy and Scale*. Island Press, Covelo, CA.

Segurado, P. & Araújo, M. B. (2004) An evaluation of methods for modelling species distributions. *Journal of Biogeography*, **31**, 1555-1568.

StatSoft (2001) *STATISTICA (data analysis software system and user´s manual). Version 6.* StatSoft, Inc., Tulsa, OK.

Teixeira, J. & Arntzen, J. W. (2002) Potential impact of climate warming on the distribution of the Golden-striped salamander, *Chioglossa lusitanica*, on the Iberian Peninsula. *Biodiversity and Conservation*, **11**, 2167-2176.

Zaniewski, A. E., Lehmann, A. & Overton, J. McC. (2002) Predicting species spatial distributions using presence-only data: a case study of native New Zealand ferns. *Ecological Modelling*, **157**, 261-280.

Zweig, M. H. & Campbell, G. (1993) Receiver-operating characteristics (ROC) plots: a fundamental evaluation tool in clinical medicine. *Clinical Chemistry*, **39**, 561-577.

# EL FANTASMA DE LOS EVENTOS NO EQUILIBRADOS EN LOS MODELOS PREDICTIVOS DE DISTRIBUCIÓN DE ESPECIES

**RESUMEN.** Las muestras en desequilibrio son consideradas un inconveniente a la hora de predecir la distribución geográfica de las especies, por lo que se ha recomendado trabajar con prevalencias de 0.5. Nosotros aquí argumentamos que las muestras en desequilibrio no son un problema desde un punto de vista estadístico, y que se pueden obtener buenos modelos siempre y cuando se elijan buenas variables predictoras y se aplique un punto de corte adecuado para convertir las probabilidades en presencia/ausencia. No deben confundirse los efectos de muestras desequilibradas con los derivados de trabajar con datos de baja calidad: presencia de falsas ausencias, bajos tamaños maestrales, o mala representación del gradiente ambiental y espacial. Finalmente, insistimos en la necesidad de invertir más esfuerzo en la mejora de la calidad de los datos de entrenamiento de los modelos y de los procesos de validación.

**Palabras clave:** modelos de distribución de especies, muestras en desequilibrio, capacidad predictiva, punto de corte

**Este capítulo ha sido publicado en:**

JIMÉNEZ-VALVERDE, A. & LOBO, J. M. (2006). The ghost of unbalanced species distribution data in geographic model predictions. *Diversity and Distributions*, en prensa.

# THE GHOST OF UNBALANCED SPECIES DISTRIBUTION DATA IN GEOGRAPHIC MODEL PREDICTIONS

**ABSTRACT.** Unbalanced samples are considered a drawback in predictive modelling of species' potential habitats, and prevalence of 0.5 has been extensively recommended. We argue that unbalanced species distribution data is not such a problem from a statistical point of view, and that good models can be obtained provided that the right predictors and cut-off to convert probabilities into presence/absence are chosen. The effects of unbalanced prevalence should not be confused with those of low-quality data affected by false absences, low sample size, or unrepresentativeness of the environmental and spatial gradient. Finally, we point out the necessity of greater research effort aimed at improving both the quality of training data sets, and the processes of validating and testing of models.

**Keywords:** species distribution models, unbalanced samples, predictive reliability, threshold cut-off

## INTRODUCTION

Species distribution modelling is now in wide use: to develop analytical and prediction tools for ecology and conservation biology (Guisan & Zimmermann, 2000; Guisan & Thuiller, 2005); to locate previously unknown populations of rare and endangered species (Raxworthy *et al.*, 2003; Guisan *et al.*, in press); to study the effect of climate warming on species distribution (Peterson, 2003; Thuiller *et al.*, 2005a), to

assess the possible impact of biological invasions (Rouget *et al*., 2004; Thuiller *et al*., 2005b), and to aid management in taking decisions (Schadt *et al.*, 2002; Barbosa *et al.*, 2003; Russell *et al.*, 2004; Chefaoui *et al.*, 2005). Quantified species-environment relationships, obtained through the development of a mathematical function linking species distribution information (usually presence/absence) to environmental predictors, are used to map decimal fraction probabilities. These probabilities are usually taken as probabilities of presence and, so, as a measure of habitat adequacy. However, probability values are highly dependent on the relative proportion of each event in the sample, being biased toward the highest number of either presences or absences, where they differ. This inherent and unavoidable bias has long been recognized by statisticians under the name of the unbalanced sample effect (Hosmer & Lemeshow, 1989). This has some important consequences for the prediction of species distributions using models and has generated confused debate in the ecological literature that is not yet resolved.

## STATISTICAL EFECTS OF UNBALANCED SAMPLES

The influence of prevalence on the performance of model predictions has repeatedly been judged to be of major importance (McPherson *et al*., 2004; Vaughan & Ormerod, 2003), leading to the supposition that the more unbalanced the samples, the less reliable the model predictions. In principle, there is no reason why the rarest events should necessarily be badly predicted, provided that models fit the data well (Cramer, 1999). Good fits can be obtained when good predictors are used and the dependent variable reflects all environmental variability. However, even in such circumstances, mean estimated probabilities of each event will be biased as a consequence of

prevalence. This bias could be especially noticeable in the case of models that do not fit the data well (Cramer, 1999), typical of those derived from field studies where the most adequate predictors are usually unknown. This interaction between model fit and prevalence bias is a question that deserves further attention.

The apparently negative effect of prevalence on prediction reliability is mediated by the cut-off value selected to convert decimal fraction probabilities to a binary variable. This cut-off should be selected appropriately to account for unbalanced samples in the conversion of the decimal fraction probabilities to presence/absence, and to evaluate the model correctly when such measures as sensitivity, specificity or the Kappa statistic, derived from a confusion matrix, are used (Fielding & Bell, 1997). As this conversion will determine model output, it will condition the cases assigned to each of the four categories of the matrix (true and false predicted presence, true and false predicted absences). The intuitively appealing 0.5 cut-off (e.g. Li *et al.*, 1997; Berg *et al.*, 2004; Meggs *et al.*, 2004) makes no sense, as each model has its own characteristics related to prevalence and fit. For example, in the case of rare species data, a 0.5 cut-off would convert presences to absences and would yield a false sensitivity value (true predicted presences) of zero in the most extreme case. In a recently published paper (Liu *et al.*, 2005), the optimum cut-off is sought through comparison of numerous criteria. Therein, the fixed 0.5 cut-off, or the widely used one which maximizes the Kappa value, were found to be among those that produced the worst results. The best presence/absence models were derived from cut-offs that maximize the sum, or minimize the difference, between sensitivity and specificity (true predicted absences), among others. Interestingly, cut-offs selected by these two criteria are highly and positively correlated with prevalence. Fig. 1 shows the relationship between prevalence and the cut-off which minimizes the difference between sensitivity and specificity (see

also Fig. 5 in Manel *et al*. 2001), using data from a simulated species and randomly resampling different training data sets varying in prevalence. Data were modelled using logistic regressions. These results suggest that the prevalence value itself could be used as a cut-off (Liu *et al*., 2005), as formerly recognized and suggested by statisticians (Cramer, 1999).

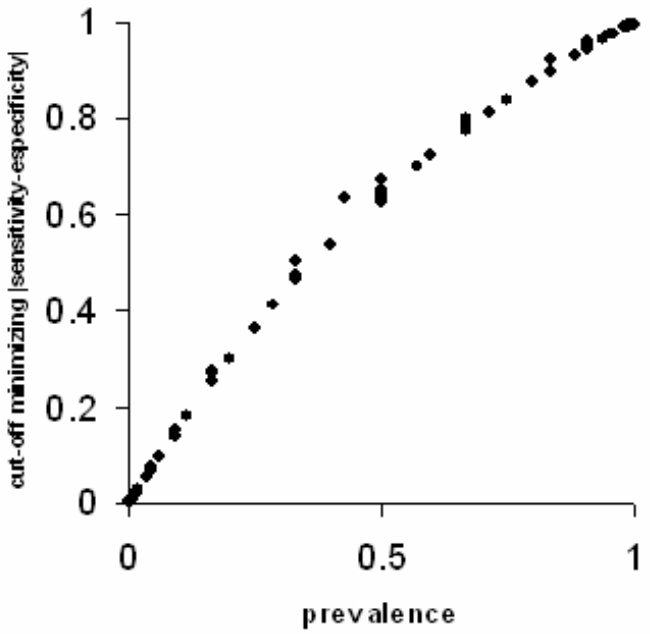



Figure 1.- Relationship between the cut-off which minimized the sensitivity-specificity difference and prevalence, using data from a simulated species and randomly resampling different training data sets varying in prevalence. Data were modelled using logistic regressions.

## CONFOUNDING FACTORS

Prevalence is a characteristic *of the data* that may sometimes correlate with species ecology, such as marginality, rarity, or specialization; these species are generally those of higher conservation concern. Bearing this in mind, caution must be exercised to avoid confusion between the effects of these biological attributes and their associated data-problems and those of prevalence.

When threshold-independent accuracy measures, such as the area under the Receiver Operating Characteristic Curve (AUC; Swets, 1988), are used to validate predictive models, confusing results have been obtained, as in some cases low prevalence values are related with high AUC scores, while the inverse has been found in other studies (see, e.g., Brotons *et al.*, 2004; Luoto *et al.*, 2005). McPherson *et al*. (2004) found best AUC scores with prevalence values around 0.5. But, if as pointed out before, there is no sound reason for models to perform poorly with unbalanced samples, what do these results mean? Effects of poor quality data can be misunderstood as false prevalence effects. For example, performance of species distribution models could depend on the sampling size of each event (independently of their relative size) and on the representativeness of the training data (i.e. presences and absences must be evenly distributed across the environmental and geographical gradient; a low sample size for an event implies poor representativeness), independently of prevalence. Additionally, the inclusion of false absences is surely a confounding factor present in many data sets whose effect will interact with prevalence. Poor quality data are usually associated with rare species, as presences are usually scarce and absences are prone to contain a high proportion of false data.

Thus, the true effect of prevalence is probably negligible when building predictive distribution models, and its "ghostly" effect is due to other puzzling factors. To avoid this supposed unbalanced-sample problem, some authors recommend resampling the training data to balance presences and absences (McPherson *et al.*, 2004; Liu *et al.*, 2005). However, in the case of reliable training data, resampling would yield only a loss of information, mainly in rare species with scarce reliable data, and should be avoided.

## RESCALING PROBABILITIES

Finally, fitted probabilities from probability maps published in research papers, if considered indicative of habitat suitability, could be misleading. While potential probability may range from 0 to 1, probabilities that do not surpass a minimum value due to low prevalence could erroneously be interpreted as low, even for well-established populations. Although it could seem paradoxical, a low value of fitted probability may be assigned to a known presence event (Pontius & Batchu, 2003), given that an under-represented event is less likely to occur in any sampling universe. To adjust the representativeness of the obtained probabilities adequately, favourability functions, such as the one proposed by Real *et al.* (in press) should be used, whose outputs are independent of prevalence due to the elimination of the random probability element. These favourability functions can be considered to be rescaling functions, as they convert logistic probabilities (*P*) into favourability values (*F*), assigning a value of *F*=0.5 to the predictor conditions for which *P*=prevalence (Real *et al.*, in press). Interestingly, whereas *P* values for different species are not comparable site to site because of the prevalence bias, this is not the case for *F* values which are directly equivalent (Real *et al.*, in press).

## COROLLARY

In conclusion, low prevalence is a property of low probability events, not a problem to be solved. Its effects on predictive tools are well known and, once accounted for, rare events should be accurately predicted if predictors are powerful and training data are reliable (especially absences) and neither spatially nor environmentally biased.

These considerations are of special relevance in conservation biology, as low prevalence is usually a property of data from endangered species. Greater research effort aimed at improving both the quality of training data sets (Vaughan & Ormerod, 2003) and the processes of validating and testing of models (Vaughan & Ormerod, 2005) should be made.

## ACKNOWLEDGMENTS

This paper was supported by a Fundación BBVA project (Diseño de una red de reservas para la protección de la Biodiversidad…) and a MEC Project (CGL2004-04309). A.J.-V. was supported by a Museo Nacional de Ciencias Naturales/C.S.I.C./Comunidad de Madrid PhD grant.

## LITERATURE CITED

Barbosa, A. M., Real, R., Olivero, J. & Vargas, J. M. (2003) Otter (*Lutra lutra*) distribution modeling at two resolution scales suited to conservation planning in the Iberian Peninsula. *Biological Conservation*, **114**, 377-387.

Berg, Å., Gärdenfors, U. & von Proschwitz, T. (2004) Logistic regression models for predicting occurrence of terrestrial molluscs in southern Sweden-importance of environmental data quality and model complexity. *Ecography*, **27**, 83-93.

Brotons, L., Thuiller, W., Araújo, M. B. & Hirzel, A. H. (2004) Presence-absence versus presence-only modelling methods for predicting bird habitat suitability. *Ecography*, **27**, 437-448.

Chefaoui, R. M., Hortal, J. & Lobo, J. M. (2005) Potential distribution modelling, niche charazterization and conservation status assessment using GIS tools: a case study of Iberian *Copris* species. *Biological Conservation*, **122**, 327-338.

Cramer, J. S. (1999) Predictive performance of binary logit model in unbalanced samples. *Journal of the Royal Statistical Society: Series D (The Statistician)*, **48**, 85-94.

Fielding, A. H., & Bell, J. F. (1997) A review of methods for the assessment of prediction errors in conservation presence/absence models. *Environmental Conservation*, **24**, 38-49.

Guisan, A., & Thuiller, W. (2005) Predicting species distribution: offering more than simple habitat models. *Ecology Letters*, **8**, 993-1009.

Guisan, A., & Zimmermann, N. E. (2000) Predictive habitat distribution models in ecology. *Ecological Modelling*, **135**, 147-186.

Guisan, A., Broennimann, O., Engler, R., Yoccoz, N. G., Vust, M., Zimmermann, N. E. & Lehmann, A. Using niche-based models to improve the sampling of rare species. *Conservation Biology*, in press.

Hosmer D.W., & Lemeshow, S. (1989) *Applied logistic regression*. Wiley, New York.
Li, W., Wang, Z., Ma, Z. & Tang, H. (1997). A regression model for the spatial distribution of red-crown crane in Yancheng Biosphere Reserve, China. *Ecological Modelling*, **103**, 115-121.

Liu, C., Berry, P. M., Dawson, T. P. & Pearson, R. G. (2005) Selecting thresholds of occurrence in the prediction of species distributions. *Ecography*, **28**, 385-393.

Luoto, M., Pöyry, J., Heikkinen, R. K. & Saarinen, K. (2005). Uncertainty of bioclimate envelope models based on the geographical distribution of species. *Global Ecology and Biogeography*, **14**, 575-584.

Manel, S., Williams, H. C. & Ormerod, S. J. (2001) Evaluating presence-absence models in ecology: the need to account for prevalence. *Journal of Applied Ecology*, **38**, 921-931.

McPherson, J. M., Jetz, W. & Rogers, D. J. (2004) The effects of species´ range sizes on the accuracy of distribution models: ecological phenomenon or statistical artefact? *Journal of Applied Ecology*, **41**, 811-823.

Meggs, J. M., Munks, S. A., Corkrey, R. & Richards, K. (2004) Development and evaluation of predictive habitat models to assist the conservation planning of a threatened lucanid beetle, *Hoplogonus simsoni*, in north-east Tasmania. *Biological Conservation*, **118**, 501-511.

Peterson, A. T. (2003) Projected climate change effects on Rocky Mountain and Great Plains birds: generalities of biodiversity consequences. *Global Change Biology*, **9**, 647-655.

Pontius, R. G. & Batchu, K. (2003) Using the relative operating characteristic to quantify certainty in prediction of location of land cover change in India. *Transactions in GIS*, **7**, 467-484.

Raxworthy, C. J., Martínez-Meyer, E., Horning, N., Nussbaum, R. A., Schbeider, G. E., Ortega-Huerta, M. A. & Peterson, A. T. (2003) Predicting distributions of known reptile species in Madagascar. *Nature*, **426**, 837-841.

Real, R., Barbosa, A. M. & Vargas, J. M. Obtaining environmental favourability functions from logistic regression. *Environmental and Ecological Statistics*, in press.

Rouget, M., Richardson, D. M., Nel, J. I., Le Maitre, D. C., Egoh, B. & Mgidi, T. (2004) Mapping the potential ranges of major plant invaders in South Africa, Lesotho and Swaziland using climatic suitability. *Diversity and Distributions*, **10**, 475-484.

Russell, K. R., Mabee, T. J. & Cole, M. B. (2004) Distribution and habitat of Columbia Torrent Salamanders at multiple spatial scales in managed forests of northwestern Oregon. *Journal of Wildlife Management*, **68**, 403-415.

Schadt, S., Revilla, E., Wiegand, T., Knauer, F., Kaczensky, P., Breitenmoser, U., Bufka, L., Cervený, J., Koubek, P., Huber, T., Stanisa, C. & Trepl, L. (2002) Assessing the suitability of central European landscapes for the reintroduction of Eurasian lynx. *Journal of Applied Ecology*, **39**, 189-203.

Swets, K. (1988) Measuring the accuracy of diagnostic systems. *Science*, **240**, 1285-1293.

Thuiller, W., Lavorel, S., Araújo, M. B., Sykes, M. T. & Prentice, I. C. (2005a) Climate change threats to plant diversity in Europe. *Proceedings of the National Academy of Sciences, USA*, **102**, 8245-8250.

Thuiller, W, Richardson, D. M., Pyšek, P., Midgley, G. F., Hughes, G. O.& Rouget, M. (2005b) Niche-based modelling as a tool for predicting the risk of alien plant invasions at a global scale. *Global Change Biology*, **11**, 2234-2250.

Vaughan, I. P. & Ormerod, S. J. (2003) Improving the quality of distribution models for conservation by addressing shortcomings in the field collection of training data. *Conservation Biology*, **17**, 1601-1611.

Vaughan, I. P. & Ormerod, S. J. (2005) The continuing challenges of testing species distribution models. *Journal of Applied Ecology*, **42**, 720-730.

# EFECTOS DE LA PREVALENCIA Y DE SU INTERACCIÓN CON EL TAMAÑO DE MUESTRA EN LOS MODELOS DE DISTRIBUCIÓN DE ESPECIES: NECESITAMOS MUCHOS MÁS DATOS DE AUSENCIA

**RESUMEN.** Aunque se considera que la prevalencia influye en la fiabilidad de los modelos predictivos de distribución de especies, poco se sabe sobre sus verdaderos efectos. En este trabajo se estudian sus consecuencias empleando una especie virtual, evitando así el posible efecto de otros factores no tenidos en cuenta (falsas ausencias, predictores sin poder explicativo, etc.) que podrían conducir a interpretaciones engañosas de los resultados. La distribución de esta especie virtual se muestreo para obtener distintos conjuntos de datos de entrenamiento con diferentes valores de prevalencia y tamaños muestrales; estos datos se modelizaron mediante regresion logística. Nuestros resultados muestran que si los predictores están relacionados con la distribución de la especie y los datos de distribución son fiables, los modelos mostrarán una elevada fiabilidad en un alto rango de prevalencias y tamaños muestrales. El efecto del tamaño de muestra aparece por debajo de 50 observaciones, y el efecto de la prevalencia lo hace a valores altos para tamaños muestrales pequeños. Sugerimos que existe una interacción entre la prevalencia, el tamaño de muestra, y la calidad de los datos de entrenaminto y de las variables predictoras. Nuestros resultados muestran la importancia de usar tantos datos de buenas ausencias como sea posible, especialmente cuando trabajemos con tamaños muestrales pequeños, típicos de especies raras y en peligro.

**Palabras clave:** regresión logística, prevalencia, tamaño de muestra, modelos predictivos de distribución, especie virtual

**Este capítulo ha sido enviado a publicar como:**

JIMÉNEZ-VALVERDE, A., LOBO, J. M. & HORTAL, J. Prevalence and its interaction with sample size effects on prediction reliability of species distribution models: we need much more absences. *Journal of Biogeography*.

# PREVALENCE AND ITS INTERACTION WITH SAMPLE SIZE EFFECTS ON PREDICTION RELIABILITY OF SPECIES DISTRIBUTION MODELS: WE NEED MUCH MORE ABSENCES

**ABSTRACT**. Even though prevalence is thought to influence the reliability of the predictions of species distribution models, little is known about its impact. Its effects were studied using a virtual species, avoiding unaccounted-for effects in the modelling process, such as false absences, non-explanatory predictors, etc. The distribution of the virtual species was subsampled to obtain several data sets of varying sample size and prevalence, and these data subsets were modelled using logistic regressions. Our results show that, providing that the predictors are truly related with the distribution of the species and that training data are reliable, models will be highly accurate over a wide variety of sample sizes and prevalence scores. The effect of sample size becomes apparent for data sets with fewer than 50 data points, and the effect of prevalence does for datasets with high prevalence values and small sample sizes. We suggest that an interaction must exist between these factors and the quality of both training data and predictor variables. Our results point out the importance of using as much as good absence data as possible, especially when dealing with small number of presences (e.g., when working with rare and endangered species).

**Key words:** logistic regression; prevalence; sample size; species distribution predictive modelling; virtual species

## INTRODUCTION

The prediction of species geographic distributions, based on known occurrences, is increasingly being used in ecology, aided by both Geographic Information Systems (GIS) and statistical quantification of species-environment relationships (Guisan & Zimmermann, 2000; Lehmann *et al.*, 2002; Rushton *et al*., 2004). Predictions based on habitat modeling techniques provide valuable data for biogeography, evolution and conservation (e.g., Peterson *et al*., 1999; Anderson *et al*., 2002a, b; Schadt *et al*., 2002; Barbosa *et al*., 2003; Peterson & Holt, 2003; Chefaoui *et al*., 2005; Jiménez-Valverde *et al*., in press).

Logistic regression (LR), commonly used to develop models derived from existing records of species distribution, predicts the probability of occurrence of an event (in this case, the presence or absence of the species) as a function of one or more independent variables. This method belongs to the so-called group discrimination techniques, methodologies that employ both presence and absence data (Guisan & Zimmermann, 2000 and Scott *et al*., 2002). Unlike profile techniques (i.e. those using only presences), group discrimination techniques take absence data into account to produce predictions and build supposedly more realistic relationships between species and environment factors (Hirzel *et al*., 2001; Brotons *et al*., 2004; Segurado & Araújo, 2004). LR, a generally robust and reliable technique that can easily be run on most software packages and implemented in a GIS environment, is also especially appropriate for the development of testable causal hypothesis (Manel *et al*. 1999a) being widely used in ecological studies (Guisan *et al.*, 2002; Lehmann *et al*., 2002; Reineking & Schröder, 2003).

Prevalence (i.e. the ratio of number of presences to total data used to build the model) is expected to have considerable effects on the estimation of model parameters and, thus, on model prediction accuracy (Vaughan & Ormerod, 2003; McPherson *et al*., 2004). However, the effect of prevalence on model accuracy remains unclear, insufficiently examined in the ecological modeling literature. Good models can be apparently obtained from low prevalence datasets, although the contrary has also been reported (e. g. Brotons *et al*., 2004; Luoto *et al.*, 2005), and others found the best models at prevalence scores around 0.5 (McPherson *et al*., 2004). To avoid these supposed negative impacts of prevalence, some authors have recommended the resampling of the data (McPherson *et al*., 2004; Liu *et al*., 2005); in fact, others have used the same number of absences and presences to fit the models, in spite of the availability of much higher numbers of absence points (e.g., Osborne *et al*., 2001; Seoane *et al*., 2006).

The main effect of prevalence is that, in the case of unbalanced samples, the probabilities derived from LR models are biased towards the highest number of either presences or absences (Hosmer & Lemeshow, 1989; Cramer, 1999). This statistical phenomenon is unavoidable, being its effect remarkably important in models with poor fit to the data (Cramer, 1999). Such effect of prevalence on model accuracy is reflected in model predictions through the selection of the LR probability cut-off used to produce the presence/absence map from the continuous probability scores. Since the probabilities attributable to the sites where the species is likely present or absent vary according to prevalence, such probability threshold determines the accuracy of the model obtained from a confusion matrix (Fielding & Bell, 1997; Manel *et al.*, 2001). Recent efforts (Liu *et al.*, 2005) aim to develop a cut-off criteria in accordance with the frequency of occurrence. However, regardless of the threshold accounting for

prevalence chosen, the effect of prevalence on model accuracy and its measurement remains open to debate. Recently, Jiménez-Valverde & Lobo (2006) suggested that prevalence has scarce effects on model accuracy, and that those supposed negative effects can be confounded with the effect of different sources of bias in poor quality data in the dependent variable, such as low sample size in the presences, lack of representation of the whole environmental gradients, or false absences. These sources of error are common in the information about most species (specially the rare ones; see, e.g., Loiselle *et al*., 2003); therefore, understanding their effects on predictive distribution modeling is of special concern in Conservation Biology.

The objective of this work is to examine the independent influence of prevalence on model accuracy, and its possible interaction with sample size, in the absence of other effects that could influence the modeling process. Since real species distributions are never completely known, it is quite difficult to assess the accuracy of distribution models and to delimit the unequivocal effects the source of uncertainty of interest. To overcome this drawback, we build a virtual species whose distribution is only conditioned by known climate variables in a simple unimodal way. Thus, by controlling the effect of the selected explanatory variables, complications due to natural conditions are avoided, allowing the identification of the genuine effects of interest.

## METHODS

**The virtual species. —** We mapped the distribution of a virtual species using actual climate data in order to use a representative scenario for the study of true species distribution patterns. Although the use of artificial species distributions to ascertain the influence of the data employed and model functions has generally been neglected in

modeling papers (but see Hirzel *et al*., 2001; Reese *et al*., 2005; Real *et al*., 2006), such virtual species are nowadays the only way to make testable experiments in predictive habitat modeling research. Using such simulations allows to:

i) ensure that LR modeling could correctly predict species distribution, avoiding the bias due to contingent, unaccounted-for or unknown explanatory factors;

ii) eliminate the random noise inherent in real biological data, and thus avoid producing overfitted models plagued by the classification errors present in real presence-absence data;

iii) provide the basis for calculating true model accuracy by comparing modeled and virtual distributions.

We mapped the distribution of the virtual species in the European region (-13º to 35º longitude, and 34º to 72º latitude), using a spatial resolution of 0.04º degrees (see Fig. 1). The total extent of the region studied was 6576.424 $km^2$ (510514 0.04º × 0.04º cells). Four environment variables (total annual precipitation, summer precipitation, mean maximum temperature and mean minimum temperature) were extracted from WORLDCLIM interpolated map database (version 1.3; see http://biogeo.berkeley.edu/worldclim/worldclim.htm). These variables were Box-Cox normalized and standardized to 0 mean and 1 standard deviation, eliminating measurement-scale effects. Principal Component Analysis (PCA, see Legendre & Legendre, 1998) was performed to obtain two reduced non-correlated environmental factors explaining 92.6 % of the environment variation across Europe, related to temperature variables (Factor 1) and to precipitation variables (Factor 2).

The distribution of the virtual species was assumed to be shaped only by these two factors. Therefore, the geographic range of the species was built using only these

two variables, so that no unknown factors affect it. The environment range inhabited by the species was set to the mean ± SD of each factor. All cells falling within these intervals for both factors were selected as the true distribution range of the virtual species in Europe (presences; *n*=91144), while the remaining cells were considered as true absences (*n*=419296; see Fig. 1). All geographic analyses were done with IDRISI Kilimanjaro software (Clark Labs, 2003).

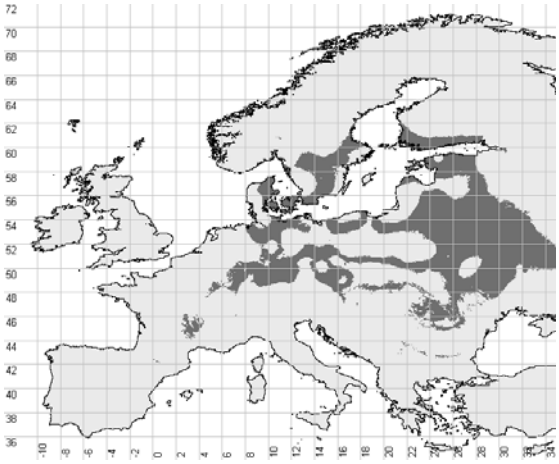


Figure 1. Spatial distribution of the virtual species in Europe. Dark grey areas are presences, and light grey absences.

**The modeling process. —** Nine different sets of presences were randomly selected from the distribution map, corresponding to increasing numbers of presence plots (*n* = 91, 456, 911, 4557, 9114, 22786, 45572, 68358 and 91144), and to successive increased percentages of presences (0.1%, 0.5%, 1%, 5%, 10%, 25%, 50%, 75% and 100%). Nine similar sets of absences were also randomly selected. All possible combinations of presences and absences were combined into presence/absence datasets (81 datasets, with *n* ranging from 182 to 182282, and prevalence ranging from 0.001 to 0.999). 14 additional datasets were used in order to detect a possible interaction with sample size (*n*=20 and 50, with 7 prevalence classes each: 0.9, 0.75, 0.6, 0.5, 0.4, 0.25, 0.1). Thus, a total of 95 datasets were modeled, varying both in the number of observations (from 20 to 182288) and in the prevalence or proportion of presences

(from 0.001 to 0.999). Since we were mainly interested in detecting the effect of prevalence, most of our samples were of a great sample size in order to avoid this possible confounding factor. Nevertheless, from the 95 models, 22 had a sample sizes lower than 1500 and 17 lower than 600.

Predictive models were based on the same environmental variables used to build the distribution of the virtual species (i.e., the two environmental factors, see above). Therefore, all variables entering in the models are truly explanatory, and no potentially explanatory factor is missing. All models were designed using logistic regression analysis (StatSoft, 2001) with a binomial error distribution. The cubic functions of the two environmental factors and their interaction term were used as explanatory variables; models were selected by a backward stepwise procedure, eliminating the non-significant terms ($p < 0.05$).

**Validation. —** The so obtained model distributions were projected onto the whole European territory. The accuracy measures used to compare true and predicted maps were derived from a confusion matrix (i.e., a cross-tabulated matrix of the number of observed presence and absence cases against the predicted presences and absences; Fielding & Bell, 1997). First of all, a cut-off was established for the logistic predictions, and all cases with predicted scores higher than such threshold were accepted as predicted presences. To do this, we calculated specificity (ratio of correctly predicted absences to the total number of absences) and sensitivity (ratio of correctly predicted presences to their total number) from the training data over a range of 100 thresholds, and selected the cut-off which minimized their difference. Such a criterion yields better results than others widely used, as it accounts for prevalence (Liu *et al*., 2005; Jiménez-Valverde & Lobo, 2006). The confusion matrix was set up after applying the threshold criterion to the model probabilities, and predicted and virtual maps were compared by

calculating sensitivity, specificity and AUC. The Receiver operating characteristic (ROC) curve is widely used as a threshold-independent accuracy measure (Zweig & Campbell, 1993; Fielding & Bell, 1997), as it seems to be the best method for model accuracy assessment (Fielding, 2002). Here, sensitivity is plotted against 1-specificity over a number of thresholds (100 in this study), and the area under the curve (AUC) calculated. AUC ranges from 0 to 1; values under 0.5 indicate discrimination worse than chance, 0.5 implies no discrimination (i.e. random predictions), and 1.0 indicates perfect discrimination.

**Testing for prevalence and sample size effects. —** In order to determine the effect of prevalence and sample size on AUC, sensitivity and specificity, scatterplots were drawn and penalized regression splines with 5 initial degrees of freedom calculated (Wood & Augustin, 2002) in order to estimate the variation explained by each independent variable. Interactions between the studied effects were also tested for significance. Splines were fitted in `R` (R Development Core Team, 2006) using the `mgcv` package (Wood, 2004). Break-points in the scatterplots were estimated by fitting regression models with segmented relationships between the dependent (accuracy measures) and independent variables (Muggeo, 2003). Segmented regressions were fitted in `R` using the `segmented` package (Muggeo, 2004).

## RESULTS

AUC scores were quite high in almost all cases (mean ± SD; 0.961 ± 0.050), being higher than 0.90 in 87 of the 95 models, and higher than 0.80 in 6 of the remaining models (Fig. 2). Only two models showed AUC scores under 0.80 (0.689 and 0.706), corresponding to the cases with smaller sample size and higher prevalence

(*n*=20 and 50; prevalence=0.9). Variations in AUC were significantly related with sample size and with the interaction between sample size and prevalence, accounting for 21.3% and 25.0% of the variability, respectively (Table 1). AUC values are consistently high until very low sample sizes are reached; segmented regressions yielded a break-point of 72.4 (Fig. 2 & 3). With 20 observations, the higher the prevalence the lower the AUC value, while with 50 observations the effect of prevalence disappears at prevalence values lower than 0.75 (see Fig. 2). When both sample size and prevalence are included together in a model in order to explain AUC variation only the interaction term is significant.

Sensitivity scores were also high and stable, with percentages of success always higher than 80% (0.929 ± 0.024) (Fig. 2). These slight variations in sensitivity were significantly related with prevalence, accounting for 11.3% of variability (Table 1); the higher the prevalence, the higher the sensitivity (Fig. 2). If low sample size cases (*n*=20 and 50) are omitted, sensitivity shows a slight increment at high prevalence values (break-point=0.99) and a decrease at low values (break-point < 0.01).

In general, specificity scores were also high (0.893 ± 0.103). Specificity scores were highly correlated with AUC values ($r = 0.99$, $p < 0.05$), so the pattern of variation with sample size and prevalence was similar for both accuracy measures (Fig. 2 & 3). The break-point for the relationship with sample size was estimated in 65.44. Variations in specificity were significantly related both with sample size and with the interaction between sample size and prevalence, accounting or 18.6% and 23.8% of variability, respectively (see Table 1). Samples with 20 observations yielded the lowest specificity values, being negatively correlated with prevalence. With 50 observations, specificity was negatively affected at prevalences higher than 0.75 (Fig. 3). Again, when both

Table 1. Effect of sample size, prevalence and interaction factors on prediction accuracy scores estimated by three statistics (columns) (* < 0.05; ** < 0.01; *** < 0.001; e.d.f, estimated degrees of freedom) tested using penalized regression splines with 5 initial degrees of freedom (Wood & Augustin, 2002).

| | **AUC** | | | **Sensitivity** | | | **Specificity** | | |
|---|---|---|---|---|---|---|---|---|---|
| | **e.d.f** | **Chi.sq** | **Exp. Dev. (%)** | **e.d.f** | **Chi.sq** | **Exp. Dev. (%)** | **e.d.f** | **Chi.sq** | **Exp. Dev. (%)** |
| Sample size | 3.52 | 24.92*** | 21.30 | 1 | 1.34 $^{ns}$ | 1.42 | 3.32 | 21.51*** | 18.60 |
| Prevalence | 1 | 2.28 $^{ns}$ | 2.39 | 1 | 11.90*** | 11.30 | 1 | 2.26 $^{ns}$ | 2.37 |
| Sample size*prevalence | 3.66 | 30.51*** | 25.00 | 1 | 0.66 $^{ns}$ | 0.70 | 3.64 | 28.59*** | 23.80 |

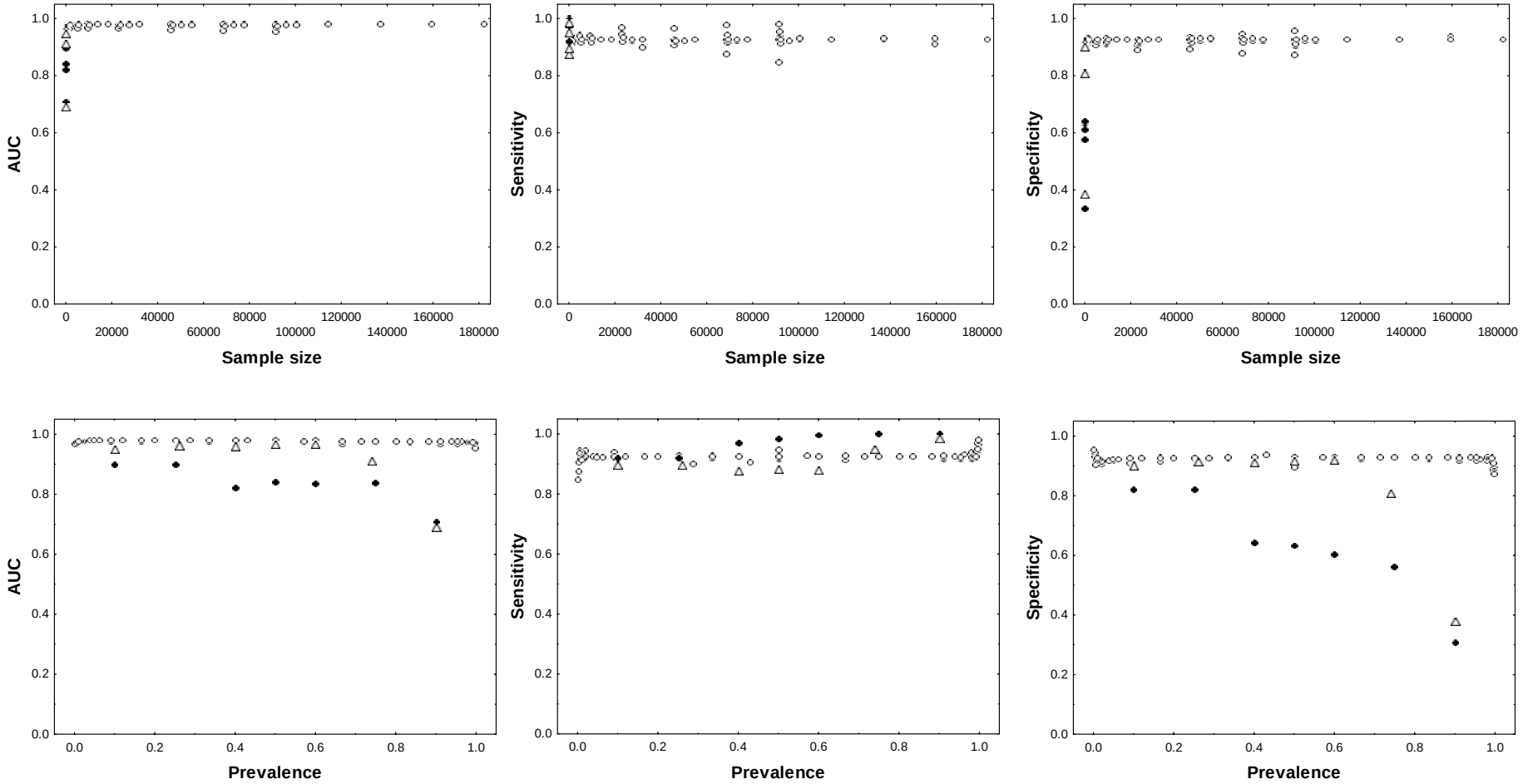



Figure 2. Relationships between the three accuracy measures (AUC, sensitivity and specificity) and sample size and prevalence (triangles, cases with *n*=50; black dots, cases with *n*=20).

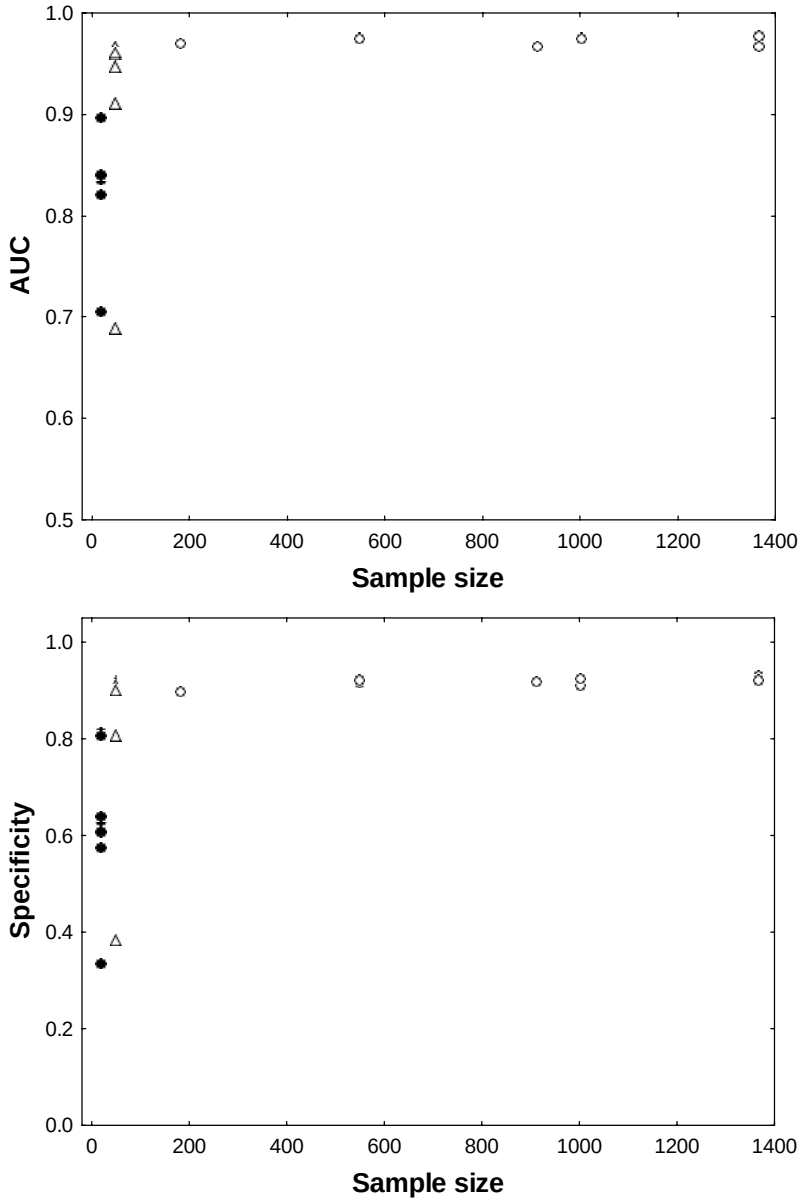



Figure 3.- Detail of the relation between sample size and AUC or specificity (triangles, cases with *n*=50; black dots, cases with *n*=20) in the interval of this last variable where performance of models starts to decrease.

sample size and prevalence are included together in a model in order to explain specificity variation only the interaction term remains significant. Specificity was also negatively correlated with sensitivity ($r$=-0.66, $p < 0.05$). If low sample size cases ($n$=20 and 50) are omitted, specificity shows a slight increment at low prevalence values (break-point < 0.01) and a decrease at high values (break-point=0.99).

# DISCUSSION

In general, quite accurate predictions were obtained from a wide range of sample sizes and prevalences. According to AUC scores, 87 predictive models were highly accurate, while six were qualified as good and useful models (following Swets, 1988). Only the two models with 20 and 50 observations and prevalences of 0.9 could be considered as having a poor discrimination capacity, according to the AUC scores. Sensitivity values were always higher than 80%, so presences are relatively well-predicted in all cases. Specificity scores are higher in almost all cases except in cases of sample size of 20 observations and prevalences higher than 0.25, as well as in the case of sample size of 50 observations and prevalences higher than 0.6. In these cases, there was a substantial overprediction.

These results are highly reliable, representing the true effects of prevalence in the controlled environment we study. Validation was done using the entire distribution, which was completely known, contrary to the real world. In real situations, modelers try to predict something that is always, and will always remain, unknown. Then, samples of the reality are used for training and validating models, and the best option for validation is to use samples as independent from the training data-set as possible. On the contrary, since we completely know the real (virtual) world, our model is compared with the whole distribution of the species. Nevertheless, we also calculated sensitivity and specificity for the data not used in the training process, and the correlations between sensitivities and specificities of the whole data set and the independent data set were 0.97 and 1.00 ($p < 0.01$), respectively. Thus, results would not have changed if validation would have been done on an independent data set.

**Prevalence effects. —** Although the effect of prevalence by itself is not important, if 20 and 50 data-point cases are not considered, prevalence affects sensitivity and specificity, reducing their values at very low and high values, respectively (see Fig. 2). Nevertheless, this effect occurs at extreme prevalence values, lower than 0.01 and higher than 0.99 and, even in those cases, sensitivity and specificity scores are higher than 0.80 and AUC scores are higher than 0.90. In any case, increasing sensitivity implies a reduction in specificity and *vice versa*. Cramer (1999) stated that there is no reason for the rarest events to be badly predicted. Prevalence affects the tests of model performance of the logistic regressions due to the mean probability biases (Fielding & Bell, 1997; Manel *et al*., 1999b; Olden *et al*., 2002). Therefore, the use of an appropriate cut-off to convert the probability map into a presence/absence one (Cramer, 1999; Liu *et al*., 1995; Jiménez-Valverde & Lobo, 2006) is central to avoid the drawbacks that could be associated to prevalence. If the frequency of occurrence is accounted for in the selection of an appropriate cut-off threshold, prevalence does not seem to have a great impact on model reliability, except in cases of extremely low and high prevalence values, where serious biases in the estimation of the parameters (King & Zeng, 2001), as well as in the reliability and coverage of the biological data (see above), may have implications in the predictive performance. Therefore, prevalence values smaller than 0.01 and higher than 0.99 should be avoided.

**Sample size effects. —** For logistic regression there is no rule in terms of minimum sample sizes (Peng *et al*., 2002). It is generally assumed that the greater the sample size, the more accurate the model (Olden & Jackson, 2000; McPherson *et al*., 2004; Martínez-Meyer, 2005; Reese *et al*., 2005). Pearce and Ferrier (2000) stated that 50 data points were not enough to obtain accurate LR models, and recommended using more than 250 points. Independently, Stockwell & Peterson (2002) stated that LR

accuracy is greatest when at least 100 data points are used; the rate of increase in accuracy was highly dependent on sample size in cases of fewer than 20 observations. In addition to this, McPherson *et al.* (2004) reported a sample size of 300 as the lower bound to obtain optimal models of South African bird distribution data. Our results seem to show that quite small sample sizes significantly reduce model prediction reliability, yielding models that tend to overpredict the virtual species distribution. Once sample size reaches a value of around 70, model reliability becomes independent of sample size.

**Confounding factors: the interaction between prevalence and sample size.** — There is a strong interaction between prevalence and sample size; this interaction means that, with small sample sizes, the higher the prevalence, the greater the distribution overprediction. Thus, it is possible to obtain moderately accurate models with sample sizes even as small as 20 data points, provided that the number of training absences is higher than the number of presences. So, why is the negative interaction of unbalanced samples not observed at low prevalences? Our virtual species can be considered "central", i. e., a species with its optimum at intermediate levels of each environment factor. Thus, while presence structures are climate dependent, absences are not, and they can be found in a greater variety of environmental situations (the most probable situation in the real world, although it depends on the extent of study, which usually is subjectively chosen). Thus, with small sample sizes and high prevalence scores it is likely that the number of absences is not enough to restrict model predictions. These results support the suggestions of Jiménez-Valverde & Lobo (2006) about the importance of the sample size of each outcome of the event (presences or absences) in order to represent the environmental gradient, independently of their relative size.

In our study the lack of spatial bias in our presences and absences, and the reliability of absence are guaranteed, which is assumed but unverifiable in real data. Sample size and its interaction with prevalence effects can be promoted by spatial aggregation and misclassification in the training data. Thus, all these facts possibly explain the differences in the minimum sample size reliable enough to build optimal models reported in the ecological literature (Pearce & Ferrier, 2000; Stockwell & Peterson, 2002; McPherson *et al.*, 2004). In addition, in real cases, apparently prevalence effects (but actually sample size of each outcome effects, see above), here only detectable with low sample sizes, may be appreciable at higher values. It seems evident that, most times, the probability of sampling spatial and environmental variation accurately increases with increasing sample size. In the same way, it is also likely that the effect of false absences (which affect model parameter estimations negatively; Tyre *et al*., 2003; Martin *et al*., 2005), would presumably diminish with increasing sample size. Thus, a good set of training data is the main factor in building powerful predictive models (Vaughan & Ormerod, 2003). Such data are only obtainable with optimal sampling protocols designed to select training points across the whole spatial and environment gradient (Wessels *et al*., 1998; Jiménez-Valverde & Lobo, 2004; Hortal & Lobo, 2005). In addition, a preliminary assessment of the reliability of absences can help in discarding those that are less credible (Anderson, 2003; Palmer *et al*., 2003).

Since confounding factors are usually unknown but may be omnipresent in the real world, the recommendation of gathering sample sizes as big as possible is probably positive (but see Stockwell & Peterson, 2002). However, this is not always affordable. In fact, sample sizes in the interval of uncertainty (< 500; Long, 1997) are more the rule than the exception in real situations. Thus, the lower the sample size, the more the relevance of well-designed sampling protocols. When working with presence data

extracted from bibliography or biological databases, then procedures to obtain pseudo-absence data make possible the selection of a great number of reliable absence data (see, for example, Ferrier & Watson, 1997; Zaniewski *et al*., 2002; Engler *et al*., 2004; Lobo *et al*., 2006; Jiménez-Valverde *et al*., in press). Reliable absences can be also obtained from places with well-known species richness inventories (see, e.g., Hortal *et al*., 2004), where the absence of a non-reported species can be considered highly certain.

Our results must be taken with caution when extrapolating to modeled species distributions in the real world. With our virtual species we have eliminated any unaccounted-for effects which are inevitably present when working with real data. Thus, historical factors, meta-population processes, interaction with other species, collinearity, unaccounted-for environment variables, or even simple stochasticity (which frequently reduces the predictive power of models) are not influencing our modeling process.

**Concluding remarks. —** Our results are consistent with the proposals of Jiménez-Valverde & Lobo (2006) about the low relevance of prevalence on the accuracy of predictive models, and the implication of other confounding factors associated which usually are correlated with prevalence. Sample size of each outcome of the event (presences or absences) and representativeness of the environmental gradient by the training data is extremely important in order to achieve reliable models. As species usually show unimodal responses to environment (Austin, 2002), the representativeness of the absence data used is especially relevant to restrict predictions and avoid high commission error rates. Situations of this kind could be found when working with rare species, for which researchers usually have a few presence points and no certain absences. In such cases, unbalanced prevalences, obtained through the creation of pseudo-absences, is a desirable property of the data instead of a scenario to

be avoided. Data resampling in order to work with prevalences of 0.5 must be discarded, since it would yield only a loss of information, especially relevant when rare species are the focus of the research.

Prevalence has been sometimes misunderstood as being a property of the species (i.e. number of grid cells with presence records/total number of cells in the region), instead of a property of the data set (i.e. number of presences used/number of observations used). This is due to the widespread practice of using all cells in the regions to model presence/absence data, without separating true absences from those due to lack of recording effort (see, e.g., Luoto *et al*., 2005). In these cases, false absences affect the results and contribute to confuse their effects with prevalence. Further investigation is needed to understand the effects of these almost always unavoidable sources or error in predictive modeling and their interactions with sample size and prevalence. The experimental approach applied in this work could help to estimate the effect of these error sources.

## ACKNOWLEDGEMENTS

This paper was supported by a Fundación BBVA project (Diseño de una red de reservas para la protección de la Biodiversidad en América del sur austral utilizando modelos predictivos de distribución con taxones hiperdiversos) and a MEC Project (CGL2004-04309). AJ-V was supported by a Museo Nacional de Ciencias Naturales/C.S.I.C./Comunidad de Madrid PhD grant. JH was supported by the Portuguese FCT (Fundação para a Ciência e Tecnologia) grant BPD/20809/2004

## LITERATURE CITED

Anderson, R. P. (2003) Real vs. artefactual absences in species distributions: tests for *Oryzomys albigularis* (Rodentia: Muridae) in Venezuela. *Journal of Biogeography*, **30,** 591-605.

Anderson, R. P., Gómez-Laverde, M. & Peterson, A. T. (2002a) Geographical distributions of spiny pocket mice in South America: insights from predictive models. *Global Ecology and Biogeography*, **11,** 131-141.

Anderson, R. P., Peterson, A. T. & Gómez-Laverde, M. (2002b) Using niche-based GIS modeling to test geographic predictions of competitive exclusion and competitive release in South American pocket mice. *Oikos*, 98**,** 3-16.

Austin, M. P. (2002) Spatial prediction of species distribution: an interface between ecological theory and statistical modelling. *Ecological Modelling*, **157,** 101-118.

Barbosa, A. M., Real, R., Olivero, J. & Vargas, J. M. (2003) Otter (*Lutra lutra*) distribution modeling at two resolution scales suited to conservation planning in the Iberian Peninsula. *Biological Conservation*, **114,** 377-387.

Brotons, L., Thuiller, W., Araújo, M. B. & Hirzel, A. H. (2004) Presence-absence versus presence-only modelling methods for predicting bird habitat suitability. *Ecography*, **27,** 437-448.

Chefaoui, R. M., Hortal, J. & Lobo, J. M. (2005) Potential distribution modelling, niche characterization and conservation status assessment using GIS tools: a case study of Iberian *Copris* species. *Biological Conservation*, **122,** 327-338.

Clark Labs (2003) *Idrisi Kilimanjaro. GIS software package*. Clark Labs, Worcester, MA.

Cramer, J. S. (1999) Predictive performance of binary logit model in unbalanced samples. *Journal of the Royal Statistical Society, Series D*, **48,** 85-94.

Engler, R., Guisan, A. & Rechsteiner, L. (2004) An improved approach for predicting the distribution of rare and endangered species from occurrence and pseudo-absence data. *Journal of Applied Ecology*, **41,** 263-274.

Ferrier, S., & Watson, G. (1997) *An evaluation of the effectiveness of environmental surrogates and modelling techniques in predicting the distribution of biological diversity*. Environment Australia, Canberra, available in http://www.deh.gov.au/biodiversity/publications/technical/surrogates/

Fielding, A. H. (2002) What are the appropiate characteristics of an accuracy measure? In *Predicting Species Occurrences. Issues of Accuracy and Scale,* eds. J. M. Scott, P. J.

Heglund, J. B. Haufler, M. Morrison, M. G. Raphael, W. B. Wall, & F. Samson, pp. 271-280. Island Press, Covelo, CA.

Fielding, A. H., & Bell, J. F. (1997) A review of methods for the assessment of prediction errors in conservation presence/absence models. *Environmental Conservation*, **24,** 38-49.

Guisan, A., & Zimmermann, N. E. (2000) Predictive habitat distribution models in ecology. *Ecological Modelling*, **135,** 147-186.

Guisan, A., Edwards, T. C. & Hastie, T. (2002) Generalized linear and generalized additive models in studies of species distributions: setting the scene. *Ecological Modelling*, **157,** 89-100.

Hirzel, A. H., Helfer, V. & Metral, F. (2001) Assessing habitat-suitability models with a virtual species. *Ecological Modelling*, **145,** 111-121.

Hortal, J., & Lobo, J. M. (2005) An ED-based protocol for optimal sampling of biodiversity. *Biodiversity and Conservation*, **14,** 2913-2947.

Hortal, J., Garcia-Pereira, P. & García-Barros, E. (2004) Butterfly species richness in mainland Portugal: Predictive models of geographic distribution patterns. *Ecography*, **27,** 68-82.

Hosmer D. W. & Lemeshow, S. (1989) *Applied Logistic Regression*. Wiley, New York.

Jiménez-Valverde, A. & Lobo, J. M. (2004) Un método sencillo para seleccionar puntos de muestreo con el objetivo de inventariar taxones hiperdiversos: el caso práctico de las familias *Araneidae* y *Thomisidae* (*Araneae*) en la Comunidad de Madrid, España. *Ecología*, **18,** 297-308.

Jiménez-Valverde, A. & Lobo, J. M. (2006) The ghost of unbalanced species distribution data in geographic model predictions. *Diversity and Distributions*, in press.

Jiménez-Valverde, A., Ortuño, V. M. & Lobo, J. M. Exploring the distribution of *Sterocorax* Ortuño, 1990 (Coleoptera, Carabidae) species in the Iberian Peninsula. *Journal of Biogeography*, in press.

King, G. & Zeng, L. (2001) Logistic regression in rare events data. *Political Analysis*, **9,** 137-163.

Legendre, P. & Legendre, L. (1998) *Numerical Ecology*. Elsevier, Amsterdam.

Lehmann, A., Overton, J. M. & Austin, M. P. (2002) Regression models for spatial prediction: their role for biodiversity and conservation. *Biodiversity and Conservation*, **11,** 2085-2092.

Liu, C., Berry, P. M., Dawson, T. P. & Pearson, R. G. (2005) Selecting thresholds of occurrence in the prediction of species distributions. *Ecography*, **28,** 385-393.

Lobo, J. M., Verdú, J. R. & Numa, C. (2006) Environmental and geographical factors affecting the Iberian distribution of flightless *Jekelius* species (Coleoptera: Geotrupidae). *Diversity and Distributions*, **12,** 179-188.

Loiselle, B. A., Howell, C. A., Graham, C. H., Goerck, C. H., Brooks, T., Smith, K. G. & Williams, P. H. (2003) Avoiding pitfalls of using species distribution models in conservation planning. *Conservation Biology*, **17,** 1591-1600.

Long, J. S. (1997) *Regression Models for Categorical and Limited Dependent Variables.* Sage Publications, Thousand Oaks, CA.

Luoto, M., Poyry, J., Heikkinen, R. K. & Saarinen, K. (2005) Uncertainty of bioclimate envelope models based on the geographical distribution of species. *Global Ecology and Biogeography*, **14,** 575-584.

Manel, S., Dias, J. M., Buckton, S. T. & Ormerod, S. J. (1999a) Comparing discriminat analysis, neural networks and logistic regression for predicting species distributions: a case study with a Himalayan river bird. *Ecological Modelling*, **120,** 337-347.

Manel, S., Dias, J. M., Buckton, S. T. & Ormerod, S. J. (1999b) Alternative methods for predicting species distributions: an illustration with Himalayan river birds. *Journal of Applied Ecology*, **36,** 734-747.

Manel, S., Williams, H. C. & Ormerod, S. J. (2001) Evaluating presence-absence models in ecology: the need to account for prevalence. *Journal of Applied Ecology*, **38,** 921-931.

Martin, T. G., Wintle, B. A., Rhodes, J. R., Kuhnert, P. M., Field, S. A., Low-Choy, S. J., Tyre, A. J. & Possingham, H. P. (2005) Zero tolerance ecology: improving ecological inference by modelling the source of zero observations. *Ecology Letters*, **8,** 1235-1246.

Martínez-Meyer, E. (2005) Climate change and biodiversity: some considerations in forecasting shifts in species´ potential distributions. *Biodiversity Informatics*, **2,** 42-55.

McPherson, J. M., Jetz, W. & Rogers, D. J. (2004) The effects of species´ range sizes on the accuracy of distribution models: ecological phenomenon or statistical artefact? *Journal of Applied Ecology*, **41,** 811-823.

Muggeo, V. M. R. (2003) Estimating regression models with unknown break-points. *Statistics in Medicine,* **22,** 3055–3071.

Muggeo, V. M. R. (2004) `segmented`: segmented relationships in regression models. R package version 0.1-4.

Olden, J. D. & Jackson, D. A. (2000) Torturing data for the sake of generality: How valid are our regression models? *Écoscience*, **7,** 501-510.

Olden, J. D., Jackson, D. A. & Peres-Neto, P. R. (2002) Predictive models of fish species distributions: A note on proper validation and chance predictions. *Transactions of the American Fisheries Society*, **131,** 329-336.

Osborne, P. E., Alonso, J. C. & Bryant, R. G. (2001) Modelling landscape-scale habitat use using GIS and remote sensing: a case study with great bustards. *Journal of Applied Ecology*, **38,** 458-471.

Palmer, M., Gómez-Pujol, L., Pons, G. X., Mateu, J. & Linde, M. (2003) Noisy data and distribution maps: the example of *Phylan semicostatus* Mulsant and Rey, 1854 (Coleoptear, Tenebrionidae) from Serra de Tramuntana (Mallorca, Western Mediterranean). *Graellsia*, **59,** 389-398.

Pearce, J. & Ferrier, S. (2000) Evaluating the predictive performance of habitat models developed using logistic regression. *Ecological Modelling,* **133,** 225-245.

Peng, C.-Y. J., Lee, K. L. & Ingersoll, G. M. (2002) An introduction to logistic regression analysis and reporting. *The Journal of Educational Research*, **96**, 3-14.

Peterson, A. T. & Holt, R. D. (2003) Niche differentiation in Mexican birds: using point occurrences to detect ecological innovation. *Ecology Letters*, **6,** 774-782.

Peterson, A. T., Soberón, J. & Sánchez-Cordero, V. (1999) Conservatism of ecological niches in evolutionary time. *Science*, **285,** 1265-1267.

R Development Core Team (2006) *R: A language and environment for statistical computing.* R Foundation for Statistical Computing, Vienna, Austria. Available at http://www.R-project.org

Real, R., Barbosa, A. M. & Vargas, J. M. (2006) Obtaining environmental favourability functions from logistic regression. *Environmental and Ecological Statistics*, in press.

Reese, G. C., Wilson, K. R., Hoeting, J. A. & Flather, C. H. (2005) Factors affecting species distribution predictions: a simulation modeling experiment. *Ecological Applications,* **15,** 554-564.

Reineking, B. & Schröder, B. (2003) Computer-intensive methods in the analysis of species-habitat relationships. In *GfÖ Arbeitskreis Theorie in der Ökologie*, ed. H. Reuter, B. Breckling & A. Mittwollen, pp. 100-117. P. Lang Verlag Frankfurt.

Rushton, S. P., Ormerod, S. J. & Kerby, G. (2004) New paradigms for modelling species distributions? *Journal of Applied Ecology*, **41,** 193-200.

Schadt, S., Revilla, E., Wiegand, T., Knauer, F., Kaczensky, P., Breitenmoser, U., Bufka, L., Červený, J., Koubek, P., Huber, T., Staniša, C. & Trepl, L. (2002) Assessing the suitability of central European landscapes for the reintroduction of Eurasian lynx. *Journal of Applied Ecology*, **39,** 189-203.

Scott, J. M., Heglund, P. J., Haufler, J. B., Morrison, M., Raphael, M. G., Wall, W. B., & Samson, F. (eds.) (2002) *Predicting Species Occurrences. Issues of Accuracy and Scale*. Island Press, Covelo, CA.

Segurado, P. & Araújo, M. B. (2004) An evaluation of methods for modelling species distributions. *Journal of Biogeography,* **31,** 1555-1568.

Seoane, J., Justribó, J. H., García, F., Retamar, J., Rabadán, C. & Atienza, J. C. (2006) Habitat-suitability modelling to assess the effects of land-use changes on Dupont´s lark *Chersophilus duponti*: A case study in the Layna Important Bird Area. *Biological Conservation*, **128,** 241-252.

StatSoft (2001) *STATISTICA (data analysis software system and user´s manual). Version 6.* StatSoft, Inc., Tulsa, OK.

Stockwell, D. R. B. & Peterson, A. T. (2002) Effects of sample size on accuracy of species distribution models. *Ecological Modelling,* **148,** 1-13.

Swets, J. A. (1988) Measuring the accuracy of diagnostic systems. *Science,* **240,** 1285-1293.

Tyre, A. J., Tenhumberg, B., Field, S. A., Niejalke, D., Paris, K. & Possingham, H. P. (2003) Improving precision and reducing bias in biological surveys by estimating false negative error rates in presence-absence data. *Ecological Applications*, **13**, 1790-1801.

Vaughan, I. P. & Ormerod, S. J. (2003) Improving the quality of distribution models for conservation by addressing shortcomings in the field collection of training data. *Conservation Biology*, **17**, 1601-1611.

Wessels, K. J., Van Jaarsveld, A. S., Grimbeek, J. D. & Van der Linde, M. J. (1998) An evaluation of the gradsect biological survey method. *Biodiversity and Conservation*, **7,** 1093-1121.

Wood, S. N. (2004) `mgcv`: GAMs with GCV smoothness estimation and GAMMs by REML/PQL. R package version 1.1-8.

Wood, S. N. & Augustin, N. H. (2002) GAMs with integrated model selection using penalized regression splines and applications to environmental modelling. *Ecological Modelling,* **157**, 157-177.

Zaniewski, A. E., Lehmann, A. & Overton, J. M. (2002) Predicting species spatial distributions using presence-only data: a case study of native New Zealand ferns. *Ecological Modelling*, **157**, 261-280.

Zweig, M. H. & Campbell, G. (1993) Receiver-operating characteristics (ROC) plots: a fundamental evaluation tool in clinical medicine. *Clinical Chemistry*, **39**, 561-577.

# EL EFECTO DE LAS FALSAS AUSENCIAS EN LOS MODELOS PREDICTIVOS DE DISTRIBUCIÓN

**RESUMEN.** El objetivo de este estudio es analizar el efecto de las falsas ausencias en los modelos predictivos de distribución. Se crearon varios conjuntos de datos a partir de una especie virtual, creando distintos niveles de falsas ausencias, distribuidas al azar y con cierta estructura especial. Mediante regresión logística se trató de modelizar la distribución de la especie usando como variables predictoras los factores usados para crearla. El funcionamiento de los modelos se evaluó tanto con los datos de entrenamiento como con la distribución real. Las predicciones de los modelos con falsas ausencias distribuidas al azar pueden ser fiables, siempre y cuando se elija bien un punto de corte a partir del cual considerar a la especie como presente. Sin embargo, si las falsas ausencias están espacialmente estructuradas, los datos de ajuste y de precisión obtenidos a partir de los datos de entrenamiento serán buenos a pesar de que los modelos serán incapaces de predecir adecuadamente la distribución real. A la luz de los resultados, invertir un mayor esfuerzo en mejorar la calidad de los datos de distribución, especialmente, de las ausencias, es necesario para producir predicciones fiables.

**Palabras clave:** modelos de distribución, falsas ausencias, precisión de los modelos, incertidumbre

**Este capítulo ha sido enviado a publicar como:**

LOBO, J. M., JIMÉNEZ-VALVERDE, A. & HORTAL, J. Effect of false absences on distribution model predictions. *Global Ecology and Biogeography*.

# EFFECT OF FALSE ABSENCES ON DISTRIBUTION MODEL PREDICTIONS

**ABSTRACT.** The aim of this study is to study the effect of false absences on species distribution model predictions. Various numbers of false absences were either randomly distributed, or spatially structured, throughout the species distribution, simulated from known environmental factors, to create a variety of training data sets. The environmental factors used to delimit the species distribution are used as explanatory in a logistic regression procedure to produce model predictions. Model prediction fit was evaluated using both training data and real species distribution. Model predictions produced from randomly distributed false absences can be accurate, depending on the threshold used to convert predicted probabilities into presences/absences. But while model predictions from spatially-structured false absences explain training-data variability acceptably, their explanation of true species distribution is inaccurate and spatially biased. Good predictions can be obtained from randomly distributed false absences by means of an appropriate cut-off threshold. However, spatially-structured (as is commonly found) false absences reduce model prediction reliability. In the light of our results, improved, true species absence data, worth additional investment, should lead to more accurate predictions from distribution maps.

**Keywords:** distribution models, false absences, model accuracy, uncertainty

## INTRODUCTION

Species distribution information is needed for both biogeographic and conservation purposes. Taking advantage of the latest developments in data processing, various initiatives aim to compile all available information (Bisby, 2000; Edwards *et al*., 2000; Godfray, 2002). However, such information generally consists of presence data only (records of species occurrence), so that species distribution maps lack any indication of true species absences. Thus, localities lacking presences may correspond to true absences (intensive sampling effort failed to find the species) or false absences (the species is present but not detected; see Anderson, 2003 or Loiselle *et al*., 2003).

Species distributions are usually modelled through statistical analysis of available presence/absence locations in environmental and/or geographic space (see, e.g., Guisan & Zimmerman, 2000). Therefore, distinguishing true from false absences is central to the reliability of species distribution hypotheses (see Ferrier & Watson, 1997; Hirzel *et al*., 2001; Zaniewski *et al*., 2002; Tyre *et al*., 2003; Engler *et al*., 2004; Gu & Swihart, 2004; Soberón & Peterson, 2005; Martin *et al*., 2005 or Lobo *et al*., 2006). This contribution explores the influence of false-absence number and distribution on distribution model predictions. As knowledge is incomplete of both the distribution range of almost all species, and of the variables affecting their distributions, a virtual species with a uni-modal response to environmental variables, with known true distribution and true determinants, was created to assess the effect of false absences. Various numbers of false absences were distributed, both at random and with spatial structure, within the distribution range of this virtual species. Logistic regression modelled the distribution of the species from these datasets, using the *a priori* established determinants of species distribution as predictors. Model predictions were

rated both for their ability to explain variability of the data used to calibrate the regression model, and for their power, i.e., their ratio of correctly predicted true-distribution presences and absences. Results are discussed in the light of current applications of model predictions to survey planning.

## METHODS

**Virtual species distribution. —** To characterize species distribution patterns representatively, real climate data was used to map the distribution of a virtual species. The study area was the West Palaearctic region (-13º to 66º longitude, and 34º to 72º latitude), with 0.04 × 0.04 degree grain size (Fig. 1). Thirty eight environmental variables, extracted from the WORLDCLIM map database (version 1.3; see http://biogeo.berkeley.edu/worldclim/worldclim.htm), were: monthly data on precipitation; mean maximum temperature and mean minimum temperature; and annual temperature and precipitation ranges. These variables were Box-Cox normalized and standardized to 0 mean and 1 standard deviation to eliminate measurement-scale effects. Principal Component Analysis (PCA, see Legendre & Legendre, 1998) yielded two uncorrelated factors explaining 88% of regional climate variability, one positively correlated with temperature, the other with precipitation variables (not shown). These two factors, of a uni-modal distribution, were assumed to completely determine the virtual species distribution in the study area. Environmental tolerance of the species was set to the mean ± SD of each factor. All cells falling within the interval of both factors were selected as constituting the true distribution range of the virtual species (presences; $n$=208501); the remaining cells were considered true absences ($n$=890356; see Fig. 1).

All geographic analyses were done with IDRISI Kilimanjaro software (Clark Labs, 2003).

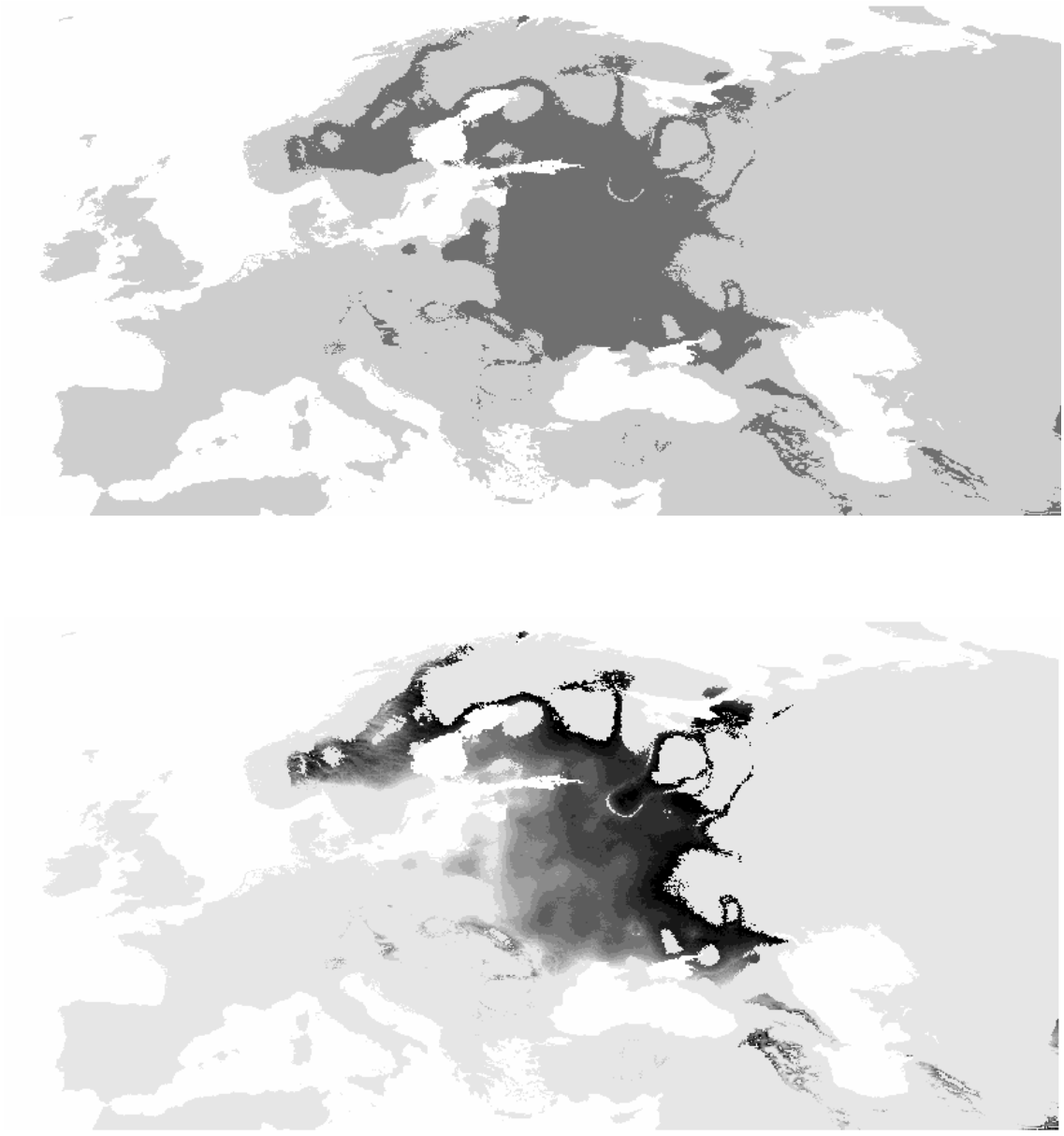


Figure 1.- Virtual species distribution (above) and map of the variation in probability of allocating false absences, used to model spatially-structured false absences (below), from 0 (light grey) to 1 (black).

**Datasets. —** Two datasets were compiled, which included all the available presence data. One of the two also included eight numbers of false absences (0, 1331, 6526, 12950, 24868, 46049, 82207 and 131744 false absences - average numbers - consecutive increases of percentage of total absence pixels from 0 % to 0.1%, 1%, 2%, 3%, 5%, 10% and 15%), randomly located within the species distribution range. False absences were selected ten times, to yield the above figures representing mean numbers of false absences distributed throughout the West Palaearctic region. The other dataset included presence data plus spatially-structured false absences; PCA scores for the first environmental factor in the species distribution range were divided into ten classes (Fig. 1), assigning a probability of allocating a false absence ten times higher to the eastern percentile of the species area (with positive PCA scores) than to the western one (with negative scores). Following this spatially structured pattern, ten numbers of false absences were randomly assigned ($n$=0, 15217, 29163, 44284, 63998, 88594, 116142, 139732, 163296 and 185277) corresponding with consecutive increases of the percentage of total absence pixels (around 0 %, 2%, 3%, 5%, 7%, 5%, 10%, 13%, 16%, 18% and 21%). Due to the low variability in the scores of the accuracy measures obtained (see Table 1) these latter datasets were modelled only once.

**Modelling. —** Species distribution was modelled by including the cubic functions of the two previously calculated environmental factors, together with their interaction terms, as explanatory variables in a backward-stepwise logistic regression analysis (StatSoft, 2001) with a binomial error distribution. Adjusted $R^2$ scores were used as a measure of explained variability. Probabilities obtained from the logistic regression were converted into presence/absence maps using a cut-off threshold. To define such a threshold, we calculated specificity (ratio of correctly predicted absences to the total number of absences) and sensitivity (ratio of correctly predicted presences to

their total number) for 100 different thresholds, selecting the cut-off that minimized their difference. All cases with predicted scores higher than such a threshold were taken as presences. Such a criterion, accounting for prevalence, seems to yield results better than those from other widely used criteria (Liu *et al*., 2005; Jiménez-Valverde & Lobo, 2006).

**Model evaluation. —** The accuracy measures used to compare observed and predicted scores were derived from confusion matrices (i.e., a cross-tabulated matrix of the number of observed presences and absences compared with predicted presences and absences; Fielding & Bell, 1997). We calculated sensitivity and specificity from the training-set matrix (the confusion matrix built with training data), as well as the frequently-used Kappa statistic, which takes into consideration both commission and omission errors and is adjusted for chance agreement (Fielding & Bell, 1997). Receiver operating characteristic (ROC) curve was also used as a threshold-independent accuracy measure (Zweig & Campbell, 1993; Fielding & Bell, 1997). Here, sensitivity is plotted against 1-specificity over a number of thresholds (100 in this study), and the area under the curve (AUC) is calculated. AUC ranges from 0 to 1; values under 0.5 indicate that the model tends to predict presences in sites where the species is absent, 0.5 implies no discrimination (i.e. random predictions), and 1.0 indicates perfect discrimination. The models were projected onto the whole study area (the West Palaearctic territory), and their probability scores converted into a binary variable (presence/absence) by applying the above-mentioned threshold criteria. Then, a new confusion matrix compared predicted and true maps by calculating the percentages of correctly predicted presences and absences, the true measure of reliability of models from regression probabilities transformed into binomial variables via threshold filtering.

## RESULTS

Increasing numbers of randomly-distributed false absences produce models with progressively lower explained variability, without effecting the accuracy of their predictions of true species distribution (Table 1; Fig. 2). As the level of false absences increases, the percentage of explained variability diminishes, at the same rate as do Kappa and AUC (Pearson correlation coefficient, $r$=0.99 and 0.98, respectively), although AUC score variability is low (Table 1). However, models with the highest proportion of false absences predict correctly a large proportion of total presences (98.1%) and a relatively large proportion of total absences (85.8%) in the training data; this implies an incorrect assignment of presences to more than 144700 points, which were absences in the training data, a figure similar to the mean number of false absences (131744). On the contrary, model predictions of the true distribution of the species are surprisingly correct: more than 98 % of presences are correctly classified, as well as 96% of absences. Thus, around 38000 absence points are incorrectly predicted as presences, and more than 70% of the total number of false absences are predicted as presences.

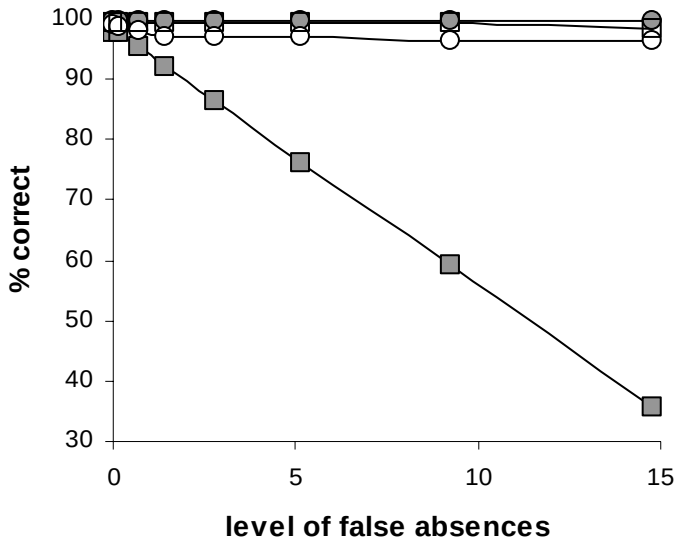



Figure 2.- Variation in the percentage of correctly predicted absences (circles) and presences (squares) with various numbers of false absences (in percentages), according to thresholds (dashed = 0.5 cut-off; open = cut-off that minimizes the difference between specificity and sensitivity; Liu *et al.*, 2005) used to convert logistic regression predictions to presences/absences. Percentages of correctly predicted absences and presences are calculated from the true virtual species distribution (see text).

Table 1.- Logistic regression model results (± SD) from various numbers of randomly distributed false absences within the virtual species distribution (Fig. 1). The last two rows correspond to probabilities obtained compared with the complete distribution of the virtual species

| **False absences** | **0** | **1331 ± 29** | **6526 ± 100** | **12950 ± 199** | **24868 ± 78** | **46049 ± 143** | **82207 ± 62** | **131744 ± 87** |
|---|---|---|---|---|---|---|---|---|
| Presences used | 208501 | 207169 ± 29 | 201974 ± 100 | 195547 ± 199 | 183633 ± 78 | 162452 ± 143 | 126294 ± 62 | 76757 ± 87 |
| Absences used | 890356 | 891688 ± 29 | 896883 ± 100 | 903229 ± 199 | 915224 ± 78 | 936405 ± 143 | 972563 ± 62 | 1022100 ± 87 |
| % explained variability | 96.61 | 95.01 ± 0.03 | 90.58 ± 0.06 | 84.10 ± 0.03 | 78.11 ± 0.03 | 69.81 ± 0.08 | 58.66 ± 0.04 | 45.89 ± 0.05 |
| AUC | 0.9997 | 0.9989 ± 0.0000 | 0.9960 ± 0.0001 | 0.9919 ± 0.0000 | 0.9854 ± 0.0001 | 0.9745 ± 0.0001 | 0.9569 ± 0.0001 | 0.9344 ± 0.0001 |
| Kappa | 0.972 | 0.9649 ± 0.0001 | 0.9311 ± 0.0001 | 0.8953 ± 0.0001 | 0.8569 ± 0.0001 | 0.7913 ± 0.0002 | 0.6720 ± 0.0001 | 0.4732 ± 0.0001 |
| **Training data** | | | | | | | | |
| % correct absences | - | 98.66 ± 0.00 | 97.21 ± 0.01 | 95.78 ± 0.01 | 94.37 ± 0.01 | 92.24 ± 0.01 | 89.21 ± 0.02 | 85.84 ± 0.01 |
| % correct presences | - | 99.45 ± 0.01 | 99.15 ± 0.03 | 99.46 ± 0.03 | 99.34 ± 0.05 | 99.49 ± 0.09 | 99.29 ± 0.15 | 98.14 ± 0.09 |
| **True distribution** | | | | | | | | |
| % correct absences | 98.98 | 98.81 ± 0.00 | 97.98 ± 0.00 | 97.15 ± 0.00 | 97.00 ± 0.00 | 96.96 ± 0.00 | 96.36 ± 0.00 | 96.28 ± 0.00 |
| % correct presences | 99.36 | 99.45 ± 0.02 | 99.41 ± 0.02 | 99.37 ± 0.03 | 99.41 ± 0.04 | 99.42 ± 0.07 | 99.18 ± 0.04 | 98.20 ± 0.05 |

Spatially-structured false absences produce the opposite effect: increasing the number of false absences does not diminish the percentage of explained variability, but does diminish model reliability (Table 2). Models with almost 8 times more false than true absences have large AUC and Kappa scores (0.997 and 0.908, respectively) and percentages higher than 97% of explained training-data variability. For the training data, only 153 presences are predicted as absences (around 0.7% of presences used), while 4410 of the absences are predicted as presences (around 0.4% of absences used). Comparison of predictions with the true distribution indicates that absences are generally well-predicted, and even if the percentage of false presences to true presences reaches 89%, only around 11900 absences are predicted as presences (1.3% of total absences). However, almost 88% of all presences were erroneously predicted as absences (183189), an underestimation of the distribution range roughly equivalent to the number of false absences used in the modelling (185277).

## DISCUSSION

**Sources of error in model predictions. —** There are three main sources of distribution model prediction error: i) quality of the dependent variable; ii) prediction power of the explanatory variables; and iii) modelling technique used to relate predictors and species data. Whilst great effort has been devoted to the design and comparison of various modelling methods, little energy has been devoted to errors coming from the first two sources.

Table 2.- Logistic regression model results from various numbers of spatially-structured false absences within the virtual species distribution (Fig. 1). Last two rows correspond to probabilities obtained compared with the complete virtual species distribution

| **False absences** | **0** | **15217** | **29163** | **44284** | **63998** | **88594** | **116142** | **139732** | **163296** | **185277** |
|---|---|---|---|---|---|---|---|---|---|---|
| Presences | 208501 | 193284 | 179338 | 164217 | 144503 | 119907 | 92359 | 68769 | 45205 | 23224 |
| Absences | 890356 | 905573 | 919519 | 934640 | 954804 | 978950 | 1006498 | 1030088 | 1053652 | 1075633 |
| % explained variability | 96.61 | 96.61 | 96.21 | 95.53 | 94.66 | 94.71 | 94.75 | 95.29 | 95.92 | 97.13 |
| AUC | 0.9997 | 0.9997 | 0.9996 | 0.9995 | 0.9993 | 0.9994 | 0.9995 | 0.9996 | 0.9997 | 0.9997 |
| Kappa | 0.9716 | 0.9674 | 0.9582 | 0.9469 | 0.9457 | 0.9345 | 0.9259 | 0.9247 | 0.9161 | 0.9079 |
| **Training data** | | | | | | | | | | |
| % correct absences | 99.10 | 98.99 | 98.79 | 98.56 | 98.64 | 98.63 | 98.78 | 99.06 | 99.29 | 99.59 |
| % correct presences | 99.19 | 99.28 | 99.06 | 98.90 | 98.96 | 99.00 | 98.97 | 99.09 | 99.25 | 99.34 |
| **True distribution** | | | | | | | | | | |
| % correct absences | 96.16 | 97.21 | 96.63 | 96.10 | 96.00 | 96.28 | 96.35 | 97.00 | 97.74 | 98.66 |
| % correct presences | 99.19 | 93.63 | 87.17 | 80.44 | 70.59 | 59.63 | 46.09 | 34.34 | 22.85 | 12.14 |

There are many different modelling techniques (see, e.g., Guisan & Thuiller, 2005) not differing essentially in their ability to identify influential variables, but rather in the complexity of the relationships established between dependent and independent variables. Techniques that establish more complex relationships between occurrence data and predictors have been found to be more accurate than techniques that do not (see Brotons *et al*., 2004; Segurado & Araújo, 2004; Elith *et al*., 2006; Araújo & Guisan, 2006 and references therein). However, regardless of the potential accuracy of these techniques, reliability of model output is highly reliant on the other two factors, which supposes major problems. For example, a modelling method may indicate significant relationships between predictors and species distribution through chance correlation, rather than causal relation. Data containing a remarkable percentage of false absences could also lead to such a result, if the selected variables model the distribution of these absences (see Olden & Jackson, 2002, and our results). Hence, along with improved strategies to estimate parameters for model predictions, it is also necessary to eliminate the bias produced by poor-quality data and irrelevant variables.

The power of predictors depends on the strength of their causal relationship with species distribution (Austin, 2002; Elith *et al*., 2002; Rushton *et al*., 2004). Identifying these predictor variables would require experimental study of the species-demography link with environmental factors. However, even if these variables are identified, other contingent or unique factors (historical factors, biotic interactions or dispersal restrictions; see Anderson *et al*., 2002; Thomas *et al*., 2004; Lobo *et al*., 2006), which are difficult to include in models, may bias results. On the other hand, the quality of the distribution data used depends on the taxonomic reliability of records; their distribution in spatial and environmental gradients (Hortal & Lobo, 2005; Reese *et al*., 2005); the number of observations (Hirzel *et al*., 2001, Stockwell & Peterson, 2002; Reese *et al*.,

2005), and the proportion of false absences in data (Moilanen, 2002; Tyre *et al*., 2003; Gu & Swihart, 2004).

**Effect of non-structured false absences and selection of most appropriate presence/absence threshold. —** The study of the effect of increasing numbers of false absences on a species with a uni-modal climate response requires an extremely well-known species distribution, shaped by known factors. As knowledge of the true geographic distribution of a species, as well as the factors that shape it, is always incomplete, we generated a virtual distribution, the direct outcome of a pair of known environmental variables. With constant effects of predictor explanatory power and of modelling techniques, and with a sample size chosen to exceed that of any limiting factor, control is assured of other sources of error, different from the number and distribution of false absences.

In our study, the number of randomly-distributed absences ends up being almost twice as large as the number of presences; the percentage of false absences varies between 0.1% and 13% of total absences. Thus, the methodological considerations derived from our results are applicable in cases of low levels of species distribution error. In such situations, model predictions can correctly classify random false absences as presences. Randomly distributed, not disproportionately numerous, false absences (noise in the description of the actual spatial response of the species to environmental gradients) would not impede logistic regressions from correctly modelling species distribution, although their explanation of training data variability may be low. In such circumstances, the effect of noise from random false absences can be eliminated by an appropriate choice of threshold to transform probabilities derived from logistic regressions into presence/absence data (see Jiménez-Valverde & Lobo, 2006).

In this paper we use a threshold that minimizes the difference between sensitivity and specificity; such prevalence-dependent threshold classifies points with relatively low probability scores as presences. This allows the correct classification of false absence plots; when false absences are not spatially structured, these plots are surrounded by presence plots, or are placed in locations with environmental conditions more similar to those of presences than true absences, showing logistic probability scores higher than those shown by true absences. Decrease in the percentage of explained variability is related to the sharing of similar environmental conditions by absence and presence points, but such an apparent decrease in model performance can be mitigated by an appropriately selected cut-off probability threshold. To demonstrate the importance of an appropriate threshold selection, we used a 0.5 cut-off (which is still being used, see e.g. Manel *et al*., 1999; Meggs *et al*., 2004; Jiménez, 2005) to map presences/absences from the previously obtained logistic probability scores. It clearly diminishes the percentage of success for the prediction of presences (Fig 2). This kind of threshold does not lead to the correct classification as presences of false absences with probability scores that, although low, are not as low as those for the locations of true absences; inaccuracy coming from such false absences remains uncorrected. The commonly employed method, selection of the threshold that maximizes Kappa scores, produces similar undesirable results (Liu *et al*., 2005; Jimenez-Valverde & Lobo, 2006). But selection of an appropriate threshold can eliminate inaccuracies caused by the incompleteness of classical distribution maps, false absences in the data, not spatially nor environmentally structured (i.e. randomly distributed).

Having said this, it is necessary to recognize that the situation simulated in this study (low proportion of false absences, and a quite high number of presences) is generally infrequent. If presences are scarce in the species distribution range, random

false absences will impede model fitting. A dramatic reduction of model prediction power could reduce the accuracy of classifications using the above-mentioned threshold selection. If noise in the data is considerable, model fit with the data will be poor, making recognition of false absences more difficult. Also in the case of unaccounted-for variables, ignored contributions to the determination of species distribution, the possibility of correct classification of false absences probably diminishes.

**Effect of spatially-structured false absences. —** Unfortunately, while proper threshold selection may eliminate model inaccuracies produced by unstructured false absences, it will not eliminate inaccuracies from non-randomly distributed false absences. In this case, both the percentage of explained variability and the scores of accuracy measures extracted from training data (AUC, Kappa, sensitivity and specificity) can be very high. However, these models are unable to adequately represent the true distribution of the species. Spatially structured false absences, more likely to be located in well-delimited environmental domains, will have logistic probability scores similar to those for areas of true absence, and any threshold will correctly classify these data. Thus, predicted presence/absence maps will unavoidably under-represent the true distribution. In our simulation, with the maximum number of false absences, both the threshold that minimizes the difference between sensitivity and specificity, and the 0.5 threshold, correctly classified only 12.1% and 10.9% of total presences, respectively. Environmental domains dominated by false absence data will consistently be identified as absences, regardless of the modelling procedure or the threshold-selection method used. In our simulation, the inclusion of false absences, spatially structured according to a geographical criterion, leads to misrepresentation of the width of species response to environmental gradients, also spatially structured. Given that such misrepresentation might be the rule, not the exception, identification of spatially (or environmentally)

structured false absences, a possible major source of model inaccuracy, may be crucial for species distribution modelling.

How are false absences distributed in currently available distribution information? The process of enlarging known species distributions by discovering new presences could be considered a sampling protocol, which is followed by taxonomists and field workers throughout time. However, such a process is likely to be spatially and/or environmentally structured (see Hortal *et al*. 2001, 2004; Lobo & Martín-Piera, 2002; Martín-Piera & Lobo, 2003). The influence of uneven sampling effort and recorder bias on available species distribution information has been thoroughly studied (Dennis *et al*., 1999; Dennis & Thomas, 2000; Zaniewski *et al*., 2002; Reutter *et al*., 2003; Graham *et al*., 2004; Martínez-Meyer, 2005), but, as far as we know, there has been no study of the possible spatial structure of distribution information from historical data accumulated by various recorders. We suspect that false absences are not randomly distributed throughout the complete distribution range of most species, especially in under-sampled regions where survey effort has been considerably lower (Cabrero-Sañudo & Lobo, 2003; see also Gaston, 2003). Our results demonstrate that apparently reliable distribution model predictions from non-randomly distributed, poor-quality data could underestimate true species distribution. Unfortunately, the current structure of survey bias, likely to produce predicted absences of a number of species in poorly-surveyed regions, seriously reduces the reliability of distribution model predictions that could otherwise improve conservation decisions there.

Our results emphasize the need to validate results with independent empirical data. Apparently accurate models (high proportion of explained variance and high scores for discrimination statistics based on training data) will fail to accurately predict species distribution if spatial bias exists in distribution information. As usually this kind

of independent data are quite difficult and costly to obtain, contrasted data-splitting validation methods have been suggested as an alternative (Fielding & Bell, 1997; Olden *et al*., 2002; Araújo & Guisan, 2006). However, if false absences are spatially structured, validation based on resubstitution methods will produce over-optimistic values of accuracy. In the same way, low explained variability scores, as well as low Kappa values, calculated from training data, do not necessarily imply faulty identification of presences and absences in the real distribution (see also Pearce & Ferrier, 2000; Fielding, 2002; Anderson *et al*., 2003; Fleishman *et al.*, 2003). Just as validation with training data can overestimate accuracy, measures of explained variation or statistics such as Kappa can underestimate model accuracy.

Tyre and collaborators (2003) demonstrate that even small numbers of false absences can have a negative impact on model performance and parameter estimates (see also Gu & Swihart, 2004). They propose using zero-inflated binomial models (where sampling units are surveyed repeatedly, up to six times) to improve the precision of estimates (see also Martín *et al*., 2005). Other authors also recommend the incorporation of detection probability in the modelling process (Royle *et al*., 2005), although information on detection probability is rarely available, especially when data do not come from standardized sampling protocols. Therefore, although presence/absence predictions can be accurate, probably robust statistical algorithms are needed to deal with false absences (Reese *et al*., 2005) if we are interested on model parameters. However, with small sample size, and a high proportion of false absences (as in most real cases), model prediction accuracy can not be unambiguously determined.

# CONCLUDING REMARKS

In recent years, advances in modelling techniques used to predict species distribution have been remarkable. However, such advances have obviated the pivotal role of dependent variable reliability, not limited to presences, but extending to absences. Our approach, the study of false absences (the most common error in species distribution data) controls some of the factors that could affect distribution model predictions. Randomly-distributed false absences can be detected through measures of explained variation. Regardless, good predictions may be obtained through the selection of an appropriate threshold to convert predicted probability scores to presences/absences. However, spatially structured false absences can yield seemingly accurate model predictions, judged by explained training-data variability, but maps of predicted presences/absences will be inaccurate and spatially biased. The only way to detect such inaccuracy is through validation with independent, reliable field data.

Good predictive models need of good species data, with correctly georeferenced and taxonomically ascribed presence plots, (Soberón & Peterson, 2004), and reliable (as Mackenzie & Royle, 2005 point out) and well-distributed absences along the whole environmental and spatial spectrum of the studied area (Hortal & Lobo, 2005). Current limitations on information about absences could be overcame by additional surveys with appropriate designs that allow to fill in the gaps in spatio-environmental coverage, but also by studying survey processes to identify true and doubtful absence plots.

## ACKNOWLEDGEMENTS

This work was supported by Spanish MEC project CGL2004-0439/BOS and Fundación BBVA project "Yamana - Diseño de una red de reservas para la protección de la biodiversidad en América del Sur Austral utilizando modelos predictivos de distribución con taxones hiperdiversos". A. J.-V. was supported by a Museo Nacional de Ciencias Naturales/C.S.I.C./Comunidad de Madrid PhD grant and JH by a Portuguese FCT (Fundação para a Ciência e Tecnologia) grant (BPD/20809/2004).

## REFERENCES

Anderson, R. P. (2003). Real vs. artefactual absences in species distributions: tests for *Oryzomys albigularis* (Rodentia: Muridae) in Venezuela. *Journal of Biogeography,* **30,** 591-605.

Anderson, R. P., Gómez-Laverde, M. & Peterson, A. T. (2002) Geographical distributions of spiny pocket mice in South America: insights from predictive models. *Global Ecology and Biogeography*, **11,** 131–141.

Anderson, R. P., Lew, D. & Peterson, A. T. (2003) Evaluating predictive models of species' distributions: criteria for selecting optimal models. *Ecological Modelling*, **162**, 211–232.

Araújo, M. B. & Guisan, A. (2006) Five (or so) challenges for species distribution modelling. *Journal of Biogeography*, in press.

Austin, M. P. (2002) Spatial prediction of species distribution: an interface between ecological theory and statistical modeling. *Ecological Modelling*, **157**, 101-118.

Bisby, F. A. (2000). The Quiet Revolution: Biodiversity Informatics and the Internet. *Science*, **289**, 2309-2312.

Brotons, L., Thuiller, W., Araújo, M. B. & Hirzel, A. H. (2004) Presence-absence versus presence-only modelling methods for predicting bird habitat suitability. *Ecography*, **27,** 437-448.

Cabrero, F. J. & Lobo, J. M. (2003) Estimating the number of species not yet described and their characteristics: the case of western Palaearctic dung beetle species (Coleoptera, Scarabaeoidea). *Biodiversity and Conservation*, **12**, 147-166.

Clark Labs. (2003) *Idrisi Kilimanjaro. GIS software package*. Clark Labs, Worcester, MA.

Dennis, R. L. H., Sparks, T. H. & Hardy, P. B. (1999) Bias in butterfly distribution maps: the effects of sampling effort. *Journal of Insect Conservation*, **3**, 33-42.

Dennis, R. L. H. & Thomas, C. D. (2000) Bias in butterfly distribution maps: the influence of hot spots and recorder's home range. *Journal of Insect Conservation*, **4**, 73-77.

Edwards, J. L., Lane, M. A. & Nielsen, E. S. (2000). Interoperability of Biodiversity Databases: Biodiversity Information on Every Desktop. *Science*, **289**, 2312-2314.

Elith, J., Burgman, M. A. & Regan, H. M. (2002) Mapping epistemic uncertainties and vague concepts in predictions of species distribution *Ecological Modelling*, **157**, 313-329.

Elith, J., Graham, C. H., Anderson, R. P., Dudík, M., Ferrier, S., Guisan, A., Hijmans, R. J., Huettmann, F., Leathwick, J. R., Lehmann, A., Li, J., Lohmann, L. G., Loiselle, B. A., Manion, G., Moritz, C., Nakamura, M., Nakazawa, Y., Overton, J. M., Peterson, A. T., Phillips, S. J., Richardson, K., Scachetti-Pereira, R., Schapire, R. E., Soberón, J., Williams, S., Wisz, M. S. & Zimmermann, N. E. (2006) Novel methods improve prediction of species' distributions from occurrence data. *Ecography*, **29**, 129-151.

Engler, R., Guisan, A. & Rechsteiner, L. (2004). An improved approach for predicting the distribution of rare and endangered species from occurrence and pseudo-absence data. *Journal of Applied Ecology*, **41**, 263–274.

Ferrier, S., & Watson, G. (1997) *An evaluation of the effectiveness of environmental surrogates and modelling techniques in predicting the distribution of biological diversity*. Environment Australia, Canberra, available in http://www.deh.gov.au/biodiversity/publications/technical/surrogates/

Fielding, A. H. (2002) What are the appropiate characteristics of an accuracy measure? In *Predicting Species Occurrences. Issues of Accuracy and Scale,* eds. J. M. Scott, P. J. Heglund, J. B. Haufler, M. Morrison, M. G. Raphael, W. B. Wall, & F. Samson, pp. 271-280. Island Press, Covelo, CA.

Fielding, A. H. & Bell, J. F. (1997) A review of methods for the assessment of prediction errors in conservation presence/absence models. *Environmental Conservation*, **24**, 38-49.

Fleishman, E., MacNally, R. & Fay, J. P. (2003) Validation Tests of Predictive Models of Butterfly Occurrence Based on Environmental Variables. *Conservation Biology*, **17**, 806-817.

Gaston, K. J. (2003) *The structure and dynamics of geographic ranges*. Oxford University Press, Oxford.

Godfray, C. (2002) Challenges for taxonomy. *Nature*, **417**, 17–19.

Graham, C. H., Ferrier, S., Huettman, F., Moritz, C. & Peterson, A. T. (2004) New developments in museum-based informatics and applications in biodiversity analysis. *Trends in Ecology and Evolution*, **19**, 497-503.

Gu, W. & Swihart, R. K. (2004) Absent or undetected? Effects of non-detection of species occurrence on wildlife–habitat models. *Biological Conservation*, **116**, 195–203.

Guisan, A. & Zimmermann, N. E. (2000) Predictive habitat distribution models in ecology. *Ecological Modelling*, **135**, 147-186.

Guisan, A. & Thuiller, W. (2005) Predicting species distribution: offering more than simple habitat models. *Ecology Letters*, **8**, 993-1009.

Hirzel, A.H., Helfer, V. & Métral, F. (2001) Assessing habitat suitability models with a virtual species. *Ecological Modelling*, **145**, 111–121.

Hortal, J. & Lobo, J.M. (2005) An ED-based protocol for optimal sampling of biodiversity. *Biodiversity and Conservation*, **14**, 2013-2947.

Hortal, J., Garcia-Pereira, P. & García-Barros, E. (2004) Butterfly species richness in mainland Portugal: Predictive models of geographic distribution patterns. *Ecography*, **27**, 68-82.

Hortal, J., Lobo, J. M. & Martín-Piera, F. (2001) Forecasting insect species richness scores in poorly surveyed territories: the case of the Portuguese dung beetles (Col. Scarabaeinae). *Biodiversity and Conservation*, **10**, 1343-1367.

Jiménez, I. (2005). Development of predictive models to explain the distribution of the West Indian manatee *Trichechus manatus* in tropical watercourses. *Biological Conservation*, **125**, 491-503.

Jiménez-Valverde, A. & Lobo, J. M. (2006) The ghost of unbalanced species distribution data in geographic model predictions. *Diversity and Distributions*, in press.

Legendre, P. & Legendre, L. (1998) *Numerical Ecology*. Elsevier, Amsterdam.

Liu, C., Berry, P. M., Dawson, T. P. & Pearson, R. G. (2005) Selecting thresholds of occurrence in the prediction of species distributions. *Ecography*, **28**: 385-393.

Lobo, J. M. & Martín-Piera, F. (2002) Searching for a predictive model for species richness of Iberian dung beetle based on spatial and environmental variables. *Conservation Biology*, **16**, 158-173.

Lobo, J. M., Verdu, J. R. & Numa, C (2006) Environmental and geographical factors affecting the Iberian distribution of flightless *Jekelius* species (Coleoptera: Geotrupidae). *Diversity and Distributions*, **12**,179-188.

Loiselle, B. A., Howell, C. A., Graham, C. H., Goerck, J. M., Brooks, T., Smith, K.G. & Williams, P. H. (2003) Avoiding pitfalls of using species distribution models in conservation planning. *Conservation Biology*, **17**, 1591-1600.

Mackenzie, D. I. & Royle, J. A. (2005) Designing occupancy studies: general advice and allocating survey effort. *Journal of Applied Ecology*, **42,** 1105–1114.

Manel, S., Dias, J. M., Buckton, S. T. & Ormerod, S. J. (1999) Comparing discriminant analysis, neural networks and logistic regression for predicting species distributions: a case study with a Himalayan river bird. *Ecological Modelling*, **120,** 337-347.

Martín-Piera, F. & Lobo, J. M. (2003) Database records as a sampling effort surrogate to predict spatial distribution of insects in either poorly or unevenly surveyed areas. *Acta Entomológica Ibérica e Macaronésica*, **1**, 23-35.

Martin, T. G., Wintle, B. A., Rhodes, J. R., Kuhnert, P. M., Field, S. A., Low-Choy, S. J., Tyre, A. J. & Possingham, H. P. (2005) Zero tolerance ecology: improving ecological inference by modelling the source of zero observations. *Ecology Letters*, **8**, 1235–1246.

Martínez-Meyer, E. (2005) Climate change and biodiversity: some considerations in forecasting shifts in species potential distributions. *Biodiversity Informatics*, **2**, 42-55.

Meggs, J. M., Munks, S. A., Corkrey, R. & Richards, K. (2004) Development and evaluation of predictive habitat models to assist the conservation planning of a threatened lucanid beetle, *Hoplogonus simsoni*, in north-east Tasmania. *Biological Conservation*, **118**, 501-511.

Moilanan, A. (2002) Implications of empirical data quality for metapopulation model parameter estimation and application. *Oikos*, **96**, 516–530.

Olden, J. D. & Jackson, D. A. (2002) A comparison of statistical models for modelling fish species distributions. *Freshwater Biology*, **47**, 1976-1995.

Olden, J. D., Jackson, D. A. & Peres-Neto, P. R. (2002) Predictive models of fish species distributions: a comment on proper validation and chance predictions. *Transactions of the American Fisheries Society*, **131**, 329-336.

Pearce, J. & Ferrier, S. (2000) Evaluating the predictive performance of habitat models developed using logistic regression. *Ecological Modelling*, **133,** 225-245.

Reese,G. C., Wilson, K. R., Hoeting, J. A. & Flather, C. H. (2005) Factors affecting species distribution predictions: a simulation model experiment. *Ecological Applications*, **15**, 554–564

Reutter, B. A., Helfer, V., Hirzel, A. H. & Vogel, P. (2003) Modelling habitat-suitability using museum collections: an example with three sympatric *Apodemus* species from the Alps. *Journal of Biogeography*, **30**, 581–590.

Royle, J. A., Nichols, J. D. & Kéry, M. (2005) Modelling occurrence and abundance of species when detection is imperfect. *Oikos*, **110**, 353-359.

Rushton, S. P., Ormerod, S. J. & Kerby, G. (2004) New paradigms for modelling species distributions? *Journal of Applied Ecology*, **41**, 193–200.

Segurado, P. & Araújo, M. B. (2004) An evaluation of methods for modelling species distributions. *Journal of Biogeography*, **31**, 1555-1569.

Soberón, J. & Peterson, A. T. (2004) Biodiversity informatics: managing and applying primary biodiversity data. *Philosophical Transactions of the Royal Society of London* B, **359**, 689–698.

Soberón, J. & Peterson, A. T. (2005) Interpretation of models of fundamental ecological niches and species' distributional areas. *Biodiversity Informatics*, **2**, 1-10.

StatSoft (2001) *STATISTICA (data analysis software system and user´s manual). Version 6.* StatSoft, Inc., Tulsa, OK.

Stockwell, D. R. B. & Peterson, A. T. (2002) Effects of sample size on accuracy of species distribution models. *Ecological Modelling*, **148**,1–13.

Thomas, C. D., Cameron, A., Green, R. E., Bakkenes, M., Beaumont, L. J., Collingham, Y. C., Erasmus, B. F. N., de Siqueira, M. F., Grainger, A., Hannah, L. Hughes, L., Huntley, B., van Jaarsveld, A. S., Midgley, G. F., Miles, L., Ortega-Huerta, M. A., Peterson, A. T., Phillips, O. L. & Williams, S. E. (2004). Extinction risk from climate change. *Nature*, **427,** 145–148.

Tyre, A. J., Tenhumberg, B., Field, S. A., Niejalke, D., Paris, K. & Possingham, H. P. (2003) Improving precision and reducing bias in biological surveys by estimating false negative error rates in presence-absence data. *Ecological Applications*, **13**, 1790–1801.

Zaniewski, A. E., Lehmann, A. & Overton, J. M. (2002) Predicting species spatial distributions using presence-only data: a case study of native New Zealand ferns. *Ecological Modelling*, **157**, 261-280.

Zweig, M. H. & Campbell, G. (1993) Receiver-operating characteristics (ROC) plots: a fundamental evaluation tool in clinical medicine. *Clinical Chemistry*, **39,** 561-577.

# DISTRIBUCIÓN POTENCIAL DE LA ARAÑA *MACROTHELE CALPEIANA* (WALCKENAER, 1805) (ARANEAE, HEXATHELIDAE) EN LA PENINSULA IBÉRICA, EXTRAPLACIÓN AL NORTE DE ÁFRICA Y A LA REGIÓN MEDITERRÁNEA, Y EVALUACIÓN DEL IMPACTO DEL CAMBIO CLIMÁTICO

**RESUMEN.** Este trabajo explora los principales determinantes de la distribución de *Macrothele calpeiana* (Walckenaer, 1805), una araña endémica de la Península Ibérica incluida en el Convenio de Berna y en la Directiva Hábitat. Se emplearon Modelos Generalizados Lineales y datos de presencia extraídos de la literatura para identificar los principales factores climáticos relacionados con la distribución de la especie, para modelizar su distribución potencial en la Península Ibérica y para extrapolar su distribución al Norte de África y a la región Mediterránea. Finalmente, se examinó el impacto del cambio climático sobre la distribución de la araña. Las variables más relevantes a la hora de delimitar el rango geográfico de la especie son aquellas relacionadas con el régimen de precipitaciones. La distribución potencial en la Península Ibérica, más amplia que la distribución actual conocida, se extiende a través de gran parte de Portugal; se propone la existencia de un factor geográfico o histórico como determinante de la ausencia de *M. calpeiana* de estas áreas adecuadas desde un punto de vista climático. La existencia de condiciones adecuadas para la araña en el Norte de África y la ausencia del género allí nos hace sugerir un origen

oriental del taxón. La extrapolación a la región Mediterránea indicó la existencia de territorio adecuado para la especie en área del Mar Egeo, donde se encuentra *Macrothele cretica* Kulczynski, 1903, la otra especie europea del género; se discute su posible origen. El calentamiento global afectará negativamente al contingente Ibérico de *M. calpeiana*, y reducirá y fragmentará el área potencial en el Norte de África. Se resalta la necesidad de confirmar la presencia o ausencia de la especie en Portugal y el Norte de África, así como de llevar a cabo estudios filogenéticos para desarrollar una hipótesis sólida sobre el origen y dispersión de *M. calpeiana*.

**Palabras clave:** cambio climático, extrapolación, Península Ibérica, *Macrothele calpeiana*, región Mediterránea, Norte de África, modelo de distribución potencial

**Este capítulo ha sido enviado a publicar como:**

JIMÉNEZ-VALVERDE, A. & LOBO, J. M. Potential Iberian Peninsula distribution of the endangered spider *Macrothele calpeiana* (Walckenaer, 1805) (Araneae, Hexathelidae), extrapolated North African and entire Mediterranean distribution, and impact of climate warming. *Diversity and Distributions*.

# POTENTIAL IBERIAN PENINSULA DISTRIBUTION OF THE ENDANGERED SPIDER *MACROTHELE CALPEIANA* (WALCKENAER, 1805) (ARANEAE, HEXATHELIDAE), EXTRAPOLATED NORTH AFRICAN AND ENTIRE MEDITERRANEAN DISTRIBUTION, AND IMPACT OF CLIMATE WARMING

**ABSTRACT.** Species distribution modelling should take all available species life-history and ecological information into account. This paper explores the main factors determining the distribution of *Macrothele calpeiana* (Walckenaer, 1805), an endemic Iberian spider included in the Bern and Habitat directives. Generalized Regression Models and presence data from the literature are used to: identify the main climate correlates within its distribution; model its potential distribution; and extrapolate that distribution to North Africa and the Mediterranean region. Finally, the impact of climate warming on the distribution of this species was also examined. Precipitation-related variables seem to be the most relevant in limiting the current geographic range of the species. Potential Iberian Peninsula distribution, much wider than at present, extends through a great part of Portugal, where the spider is not now found. A geographical or historical factor is proposed as a contributor to the absence of *M. calpeiana* from these suitable areas. Existence of suitable conditions for the species in North Africa and the absence of the genus there favors a hypothesis of oriental origin. Extrapolation to the Mediterranean region highlighted suitable territory in the Aegean

area, where *Macrothele cretica* Kulczynski, 1903, the other European *Macrothele* species, is found; its possible origin is discussed. Climate warming will negatively affect the existing Iberian *M. calpeiana* population, and will reduce and fragment potential North African habitat. Confirmation of species presence in, or absence from, Portugal and North Africa is highlighted as necessary, as well as the development of phylogenetic studies to establish a solid hypothesis of *M. calpeiana* origin and dispersion history.

**Keywords:** climate warming, extrapolation, Iberian Peninsula, *Macrothele calpeiana*, Mediterranean region, North Africa, potential distribution model

## INTRODUCTION

The genus *Macrothele* Aussserer, 1871, composed of 26 species distributed from Western Europe to Japan, belongs to the family Hexathelidae, a spider family of Gondwanic origin (Raven, 1980). The family is composed of 11 genus, the bulk occurring in Australia, New Zealand and Tasmania (Platnick, 2006). Only two, *Mediothele* Raven & Platnick, 1978 and *Scotinoecus* Simon, 1892, are endemic to South America (Platnick, 2006). As occurs with some Asian species, the taxonomic status of three *Macrothele* species, exclusively central African (Platnick, 2006), is uncertain due to lack of proper descriptions and known males (Snazell & Allison, 1989); one species was described in 1903 and the other two in 1965, with no new data since. Descriptions of the only two known European species, *M. cretica* Kulczynski, 1903 endemic to the island of Crete, and *M. calpeiana* (Walckenaer, 1805) from the

Iberian Peninsula, are now up-to-date and precise (see Blasco & Ferrández, 1986 and Snazell & Allison, 1989).

Since first published in Blasco & Ferrández (1986), new *M. calpeiana* locations in the Iberian Peninsula have slowly been outlining its distribution. Currently, *M. calpeiana* is known to be distributed in six apparently isolated core Iberian areas (see Fig. 1). Snazell (1986) and Snazell & Allison (1989) stated that the distribution of *M. calpeiana,* in accordance with this distributional pattern, corresponds with areas of high precipitation, warm winters and high summer temperatures. Of the two North African records, the first, published in 1849 from El Arrouch, Algeria, is doubtful due to the deterioration of the specimen (Lucas, 1849); the second, from Ceuta, has recently been confirmed (Ferrández & Fernández de Céspedes, 1996). These African records have led some authors to hypothesize the possible existence of this spider species in North Africa (Helsdingen & Decae, 1992), and to suggest that *M. calpeiana* could have colonized the Iberian Peninsula from there in recent times. However, Ferrández & Fernández de Céspedes (1996) proposed that, whether or not the species is of African origin, given the anthropogenic habitat of the record from Ceuta, this particular population may be an intruder. They also suggested that reduction of a past, wider *M. calpeiana* Iberian distribution could have produced the present-day apparent isolation of populations.

Recent developments in GIS data and techniques, as well as in statistical tools, enable quantification of species-environment relationships and prediction of species geographic distribution from confirmed occurrences (Guisan & Zimmermann, 2000; Scott *et al.*, 2002). Existing distribution maps are improved (Bustamante & Seoane, 2004) by such prediction modelling, while their automatic fitting procedures even outperformed models based on expert opinion (Seoane *et al*., 2005). Habitat

requirement modelling has been widely employed, among many other uses (see Guisan & Thuiller, 2005 for an overview), to: focus sampling effort aimed at locating new species or populations (e. g. Raxworthy *et al.*, 2003; Guisan *et al.*, 2006); explore biogeographic questions (Gallego *et al*., 2004; Anderson *et al*., 2002; Lobo *et al*., 2005; Jiménez-Valverde *et al*., in press); develop management decisions and conservation strategies (Schadt *et al*., 2002; Barbosa *et al*., 2003; Hirzel *et al*., 2004; Russell *et al*., 2004).

In recent times, model predictions have also been used to study the effects of climate warming on species distribution (Peterson *et al*., 2002; Peterson, 2003; Thuiller *et al*., 2005a). There is broad agreement among researchers on a direct climate-warming link with greenhouse gas emissions (Oreskes, 2004; King, 2005). Global temperatures have increased by 0.6ºC over the last century, and are predicted to increase from 1.4 - 5.8 ºC by 2100 (IPCC, 2001). Climate change will modify many environmental factors that determine species distribution and abundance (Hulme, 2005; Lovejoy & Hannah, 2005) and, nowadays, undeniable evidence of the impact of climate warming on species range shifts has appeared (Parmesan & Yohe, 2003; Root *et al.*, 2003; Hickling *et al*., 2006). Estimates of potential extinction risks (Thomas *et al*., 2004; Malcom *et al*., 2006) point to the seriousness of climate warming, probably leading to habitat loss, of special concern in the case of species with narrow geographic ranges (Thuiller *et al*., 2005b).

Although *M. calpeiana* is included in the Bern and Habitat directives, little is known about the factors that determine its distribution, and quantitative analysis is completely lacking. Thus, the objectives of this study are to: i) identify major climate correlates of *M. calpeiana* distribution in the Iberian Peninsula; ii) elaborate a hypothesis for distribution of the spider in the Iberian Peninsula; iii) identify suitable

environment for potential species distribution in both North Africa and the entire Mediterranean region through extrapolation of model predictions; and iv) predict the impact of climate warming on spider distribution in both the Iberian Peninsula and potentially suitable North Africa.

## METHODS

**Biological data. —** *M. calpeiana* presences (92) in the Iberian Peninsula, extracted from the literature (see Annex), were located in 100 km$^2$ UTM squares (Fig. 1). As the appearance of the species is quite distinctive, all presences are considered reliable.

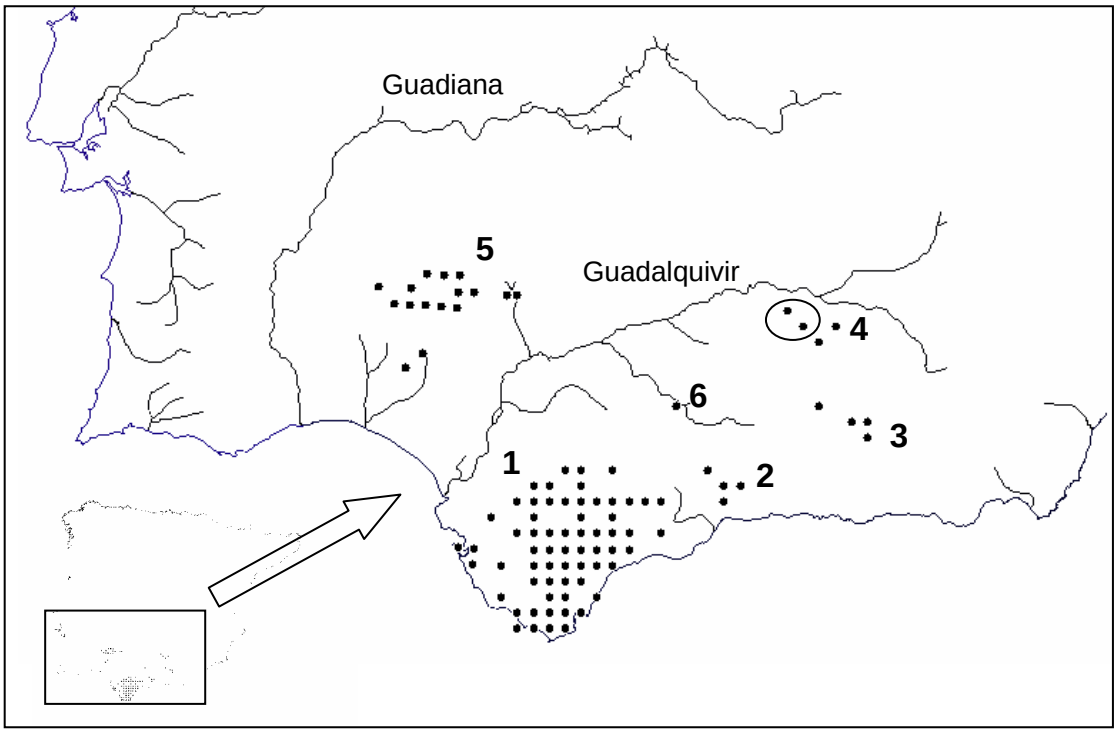



Figure 1.- Distribution of *Macrothele calpeiana* (Walckenaer, 1805) in the Iberian Peninsula. Dots represent known 100 km$^2$ UTM squares. Main rivers are shown. The circle marks the two presence points wrongly predicted by the model.

**Environmental data. —** Seven potential explanatory environment variables, selected to reflect most climate variation in each 100 $km^2$ Iberian Peninsula UTM square are: average monthly mean temperature (*temp*); average monthly maximum temperature (*tmax*); average monthly minimum temperature (*tmin*); average monthly precipitation (*precp*); seasonal precipitation variation (coefficient of variation, *precpsea*); precipitation in the wettest month (*precpw*); and precipitation in the driest month (*precpd*). All variables were obtained from WORLDCLIM interpolated map database, version 1.3 (http://biogeo.berkeley.edu/worldclim/worldclim.htm; see Hijmans *et al*., 2005).

**Creating pseudo-absences. —** As only *M. calpeiana* presence data is reliable, pseudo-absences were created for modelling the potential distribution. Although there are methods for modelling potential species distribution from presence-only data, these methods tend to overestimate distributions, because of the lack of true absences that would constrain predictions where needed (Ferrier & Watson, 1997; Zaniewski *et al*., 2002; Engler *et al*., 2004). To reduce such overestimation, use of pseudo-absences has been encouraged (Engler *et al*., 2004; Lobo *et al*., 2006; Jiménez-Valverde *et al*., in press).

Probable absences (828 pseudo-absences, giving a prevalence = 0.1), to be used in subsequent modelling, were selected at random from the area falling outside the envelope defined by the 7 environment variable maximums and minimums within the observed presence localities. This so-called niche-based envelope model (see Busby, 1986 and Beaumont *et al*., 2005) is made up of those environmental conditions, at presence points, suitable for the species.

**Modelling process. —** Among existing modelling methods (see, for example, Guisan, & Zimmermann, 2000), regression methods are probably the most popular, due to ease of implementation and interpretation of their results (Lehmann *et al.*, 2002a). Correctly applied, they perform relatively well with a training data set free of major errors (Manel *et al*., 1999a, b; Moisen & Frescino, 2002; Elith et al., 2006). This study used two regression methods, Generalized Additive Models (GAMs) and Generalized Linear Models (GLMs), the former to explore and the latter to predict.

We applied GAMs (Hastie & Tibshirani, 1990) with a logit link function to explore the relationship between *M. calpeiana* presence/pseudo-absence data and the seven selected environment variables. GAMs are semi-parametric extensions of GLMs, dealing with non-linear and non-monotonic relationships by applying smoothing functions to each predictor (Guisan *et al*., 2002; Lehmann *et al*., 2002b). In other words, GAMs can be applied to more complex response shapes than GLMs and, for a given number of degrees of freedom, fit data more closely (Wintle *et al.*, 2005). This makes GAMs ideal for the exploration of the shape of the response of spider data to each explanatory variable, to subsequently define variable relationship for GLMs model parametrization (see, for example, Olivier & Wotherspoon, 2005). GAMs with penalized regression splines were used, where the smoothing parameter is estimated by using the Un-Biased Risk Estimator criterion (UBRE), which is an approximation to the Akaike Information Criterion (AIC; Wood, 2000; Wood & Augustin, 2002). Smoothed terms with 4 initial degrees of freedom were regressed against the response variable. GAMs were fitted in `R` (R Development Core Team, 2004) using the `mgcv` package (Wood, 2004).

Multi-collinearity of independent variables negatively affects automated stepwise variable selection in regression analysis (Feinstein, 1996). To reduce the

number of possibly correlated factors, Pearson´s correlation coefficient (*r*) was calculated for variables, and groups of correlated factors were defined using an *r* value ≥ 0.8 (Silva & Barroso, 2004); variables were replaced by just one factor per group. Factors explaining the smallest deviance in GAMs, or relating in a complex or unrealistic way with presence-absence data, were dropped.

GLMs (McCullagh & Nelder, 1997) were then used to fit the environmental model to the factor selected. GLMs are extensions of linear regressions, able to deal with non-Gaussian probability distributions. Logistic regressions (GLMs with binomial distribution and logit link function) were selected for the prediction step, as they are less prone to overfitting than GAMs, and so are more likely to produce more reliable generalizations and extrapolations (Reineking & Schröder, 2003; see, for a case example, Olivier & Wotherspoon, 2005). Variables selected in the previous step were introduced in the model and further selected by a backward-stepwise procedure (Harrell, 2001). Nested models were tested using AIC criterion (Buckland *et al*., 1997), a method that penalized the log-likelihood of the model as function of the number of degrees of freedom. GLMs were fitted in R (R Development Core Team, 2004).

Autocorrelation of residuals was examined to detect possible spatial patterns in the residuals of the prediction function, and if so, any major unaccounted-for variable. First of all, residuals were calculated after correcting for the prevalence bias in logistic probabilities (see below). Moran´s *I* coefficient, which describes the degree of spatial autocorrelation for distinct distance classes, was employed with a lag distance of 30 kilometers. Moran's *I* test was checked for significance with the Bonferroni-corrected significance level (Sawada, 1999).

**Evaluation. —** Accuracy in model predictions was assessed using a jackknife procedure, a technique which yields relatively unbiased estimates of model performance (see Olden *et al*., 2002). This procedure excludes one observation, and then the model is parametrized again with the remaining *n*-1 observations to obtain a predicted probability for the excluded observation. This procedure is repeated *n* times (one per observation), and the receiver operating characteristics (ROC) technique is applied, using the area under the curve (AUC) as a measure of overall accuracy (Fielding & Bell, 1997; Pearce & Ferrier, 2000). In addition, sensitivity and specificity (presences correctly predicted as presences, and absences correctly predicted as absences, respectively) were also calculated from these new jackknife probabilities. As these two accuracy measures depend on a threshold value, above which probabilities are considered as presences, we applied the threshold which minimizes the difference between sensitivity and specificity (MDT threshold; Jiménez-Valverde & Lobo, 2006). All validation computations were run in R (R Development Core Team, 2004).

**Extrapolation of models. —** Niche models were projected on North Africa and the entire Mediterranean region. As extrapolation of model predictions beyond the range of values of the variables used in parametrization is quite unreliable, areas within these ranges were first defined and extrapolation limited to them. To asses the impact of climate warming, models were projected onto a future Iberian Peninsula and North African climate dataset (CCM3 climate model for $2100_{AD}$), predicting a doubling of $CO_2$ concentration (see Govindasamy *et al*., 2003 for details). Variables for these extrapolation scenarios were obtained from WORLDCLIM (http://biogeo.berkeley.edu/worldclim/worldclim.htm).

**Map representations. —** Probabilities derived from logistic regressions are inevitably affected by prevalence; mean probabilities are biased towards the most common outcome (Cramer, 1999). Thus, these probabilities cannot be considered indicative of habitat suitability, and they must be rescaled (Jiménez-Valverde & Lobo, 2006). For this purpose, we have used the favorability function developed by Real *et al*. (in press) (see also Muñoz *et al*., 2005) to eliminate the random element from the logistic equation:

$$F = 1 - \frac{1}{1 + e^{\left(\ln\frac{P}{1-P} - \ln\frac{n_1}{n_0}\right)}}$$

where $F$ is the favorability value, $P$ is the logistic probability, $n_1$ is the number of presences and $n_0$ is the number of absences.

## RESULTS

Relationships of the seven variables with *M. calpeiana* presence/pseudo-absence data were statistically significant (Fig. 2). Precipitation-related variables explained the greatest amount of deviance. Seasonal precipitation variation is the most important variable (66.2% of explained deviance), with a positive linear relationship. Precipitation in the driest month is the second variable in relevance, explaining a bit less than the former (52.8%), and negatively, linearly related with spider presence-absence. Average monthly precipitation and precipitation in the wettest month explain 37.7% and 26.4% of deviance, respectively; relations of both with spider presence (maximum probability of presence at ~800 mm. and ~125 mm., respectively) were

bell-shaped. Average monthly mean and minimum temperature relations with *M. calpeiana* presence-absence were positive, quasi-linear, explaining 26.8% and 24.9% of deviance, respectively. Average monthly maximum temperature is the variable that explains the least deviance (14.6%), probability of presence is positively, linearly related with this variable below ~25ºC, above which the probability reaches an asymptote.

Correlation analysis identified three pairs of highly-correlated variables ($r \geq 0.8$): average monthly precipitation-precipitation in the wettest month; precipitation in the driest month-average monthly maximum temperature; and average monthly mean temperature-average monthly minimum temperature. From the first pair, wettest-month precipitation was discarded as it explained less deviance than its partner. From the second pair, even though it explained less variation than its partner, maximum temperature was selected, since its non-linear relation with *M. calpeiana* suggests better characterization of spider environmental requirements. Additionally, driest-month precipitation is highly correlated with seasonal variation in precipitation ($r$=-0.77). From the third pair, minimum temperature was selected, as its relation with probability of presence was simpler than that of mean annual temperature, while deviance explained was similar.

As a result of the above-mentioned analysis, the following variables were introduced in the GLM: seasonal variation in precipitation (linearly related with dependent variable), average monthly precipitation (a second-term polynomial), average monthly maximum temperature (second-term polynomial) and average monthly minimum temperature (linear). The final GLM environmental model retained all variables (Table 1), accounting for 89.98% of deviance, and yielded accuracy measures greater than 0.96. Positive and significant autocorrelation scores, present in

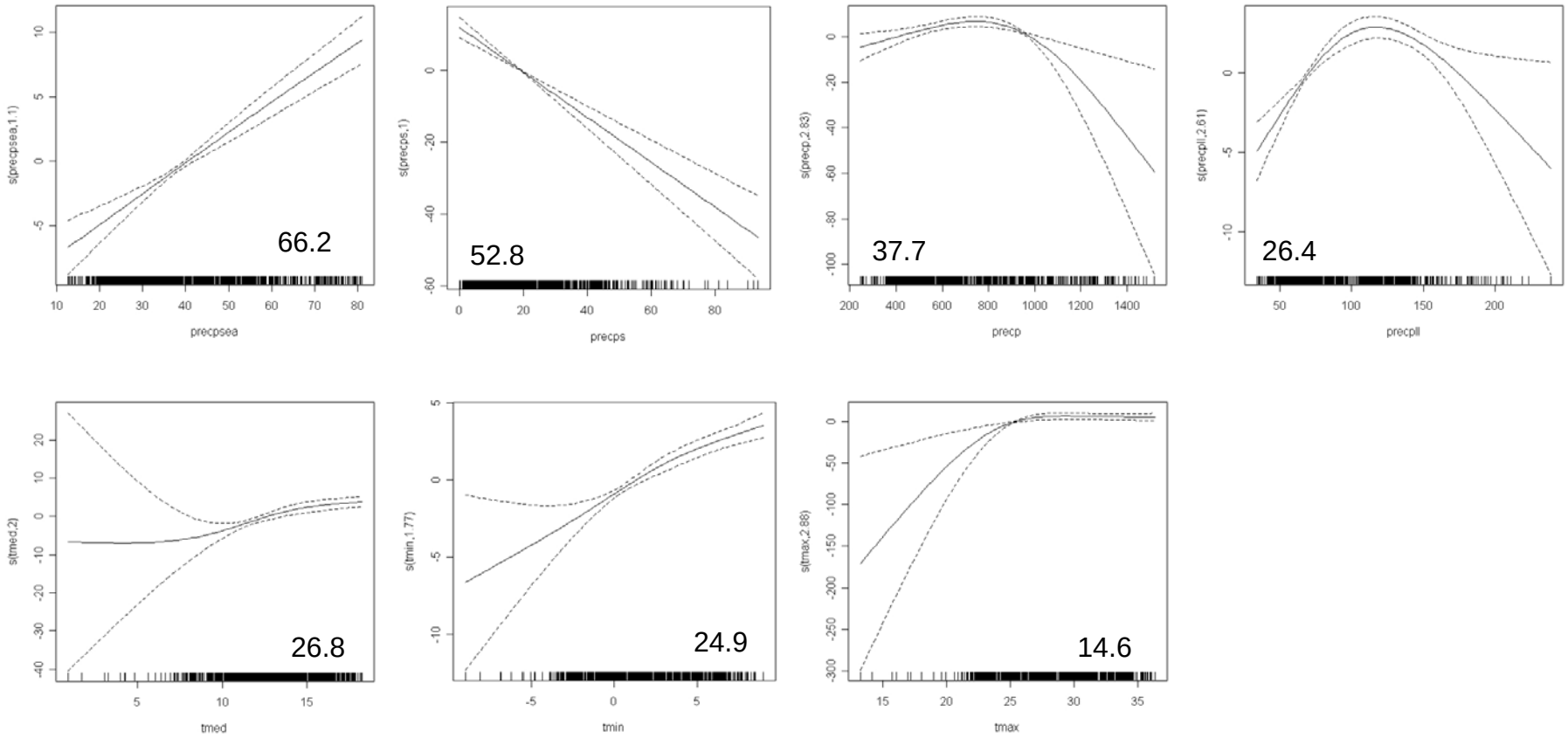



Figure 2.- Estimated GAM terms describing the relationships of *Macrothele calpeiana* (Walckenaer, 1805) with the seven environmental predictors. Estimates are shown as solid lines, 95% confidence intervals as dashed lines and cases as a rough plot along the bottom of the graphs. Explained deviance is shown as a percentage. All predictors showed *p*-values lower than 0.001 (Chi. sq. test).

Table 1.- Summary of results of the GLM model. *D*, explained deviance ([null deviance-residual deviance]/null deviance*100); AUC, area under the ROC curve; *sens*, sensitivity (proportion of presences predicted as presences); *spec*, specificity (proportion of absences predicted as absences) (* <0.05 ***<0.001).

| Terms | Estimate | Std. Error | z value |
|---|---|---|---|
| Intercept | -38.32 | 8.23 | -4.65*** |
| *tmax* | 810.69 | 209.45 | 3.87*** |
| $tmax^2$ | -605.21 | 141.78 | -4.27*** |
| *tmin* | -0.97 | 0.46 | -2.10* |
| *precp* | -337.91 | 155.77 | -2.17* |
| $precp^2$ | -632.58 | 162.50 | -3.89*** |
| *precpsea* | 0.50 | 0.14 | 3.50*** |
| **Measures of fit and predictive accuracy** | | | |
| *D*=89.98 | AUC=0.99 | *Sens*=0.97 | *Spec*=0.97 |

the two first distance lags (Fig. 3), persist even though such spatial terms as the third degree polynomial term of latitude and longitude are added (Legendre & Legendre, 1998); addition of these spatial variables does not increase deviance explained by the final climate model.

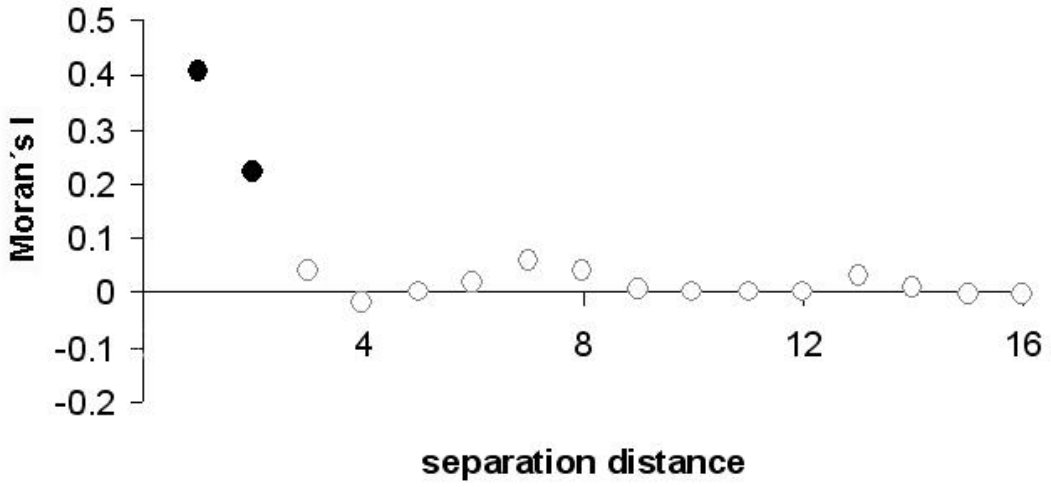



Figure 3.- Correlogram for the residuals of the predictive model calculated after rescaling the probabilities with the favorability function developed by Real *et al*. (in press) (see text for details). The lag distance is 30 kilometers and Moran's *I* autocorrelation scores were checked for significance (black dots are statistically significant) with a Bonferroni-corrected significance level (Sawada, 1999).

Figure 4 illustrates the importance of accounting for prevalence in the mapping of probabilities as an index of habitat suitability, and in the selection of a cut-off to convert probabilities into presence-absence maps. Figure 4A shows the logistic probabilities predicted for the entire Iberian territory. The MDT cut-off yielded the presence-absence map of Figure 4B. The classic 0.5 threshold would yield the map of Fig. 4C, which clearly differs from the previous one. Application of the favorability function produced the map of Figure 4D, which can be considered a habitat suitability map due to the elimination of the random element of logistic regression. As expected, application of the 0.5 cut-off to these favorability values produces a presence-absence map virtually identical to the one produced using the MDT criterion. In all the following representations, favorability scores will be used.

In the Iberian Peninsula, the model predicts broader potential *M. calpeiana* distribution than that presently known (Fig. 4); apparently isolated core areas are joined and predicted distribution extends halfway up the Iberian Peninsula and through southern and mid- Portugal, touching the Sierras of Nogueira and Mogadouro in the north. Potential climate distribution does not include the Guadalquivir Valley, except around the river mouth.

Extrapolation to North Africa (Fig. 5) identifies environmentally suitable areas in: Morocco, through the Tangier Peninsula, the Rif and most of the Atlas Mountain Range; north of Algeria and Tunisia; and two small areas in the north of Libya. Extrapolation to the Mediterranean region (Fig. 6), within model parameter environmental limits, identifies potential distribution area in: Sardinia and Sicily; some southern areas of Italy and Greece; Crete, most of the Aegean islands and Cyprus; and a large part of Turkey, Syria, the Lebanon, Israel and Jordan.

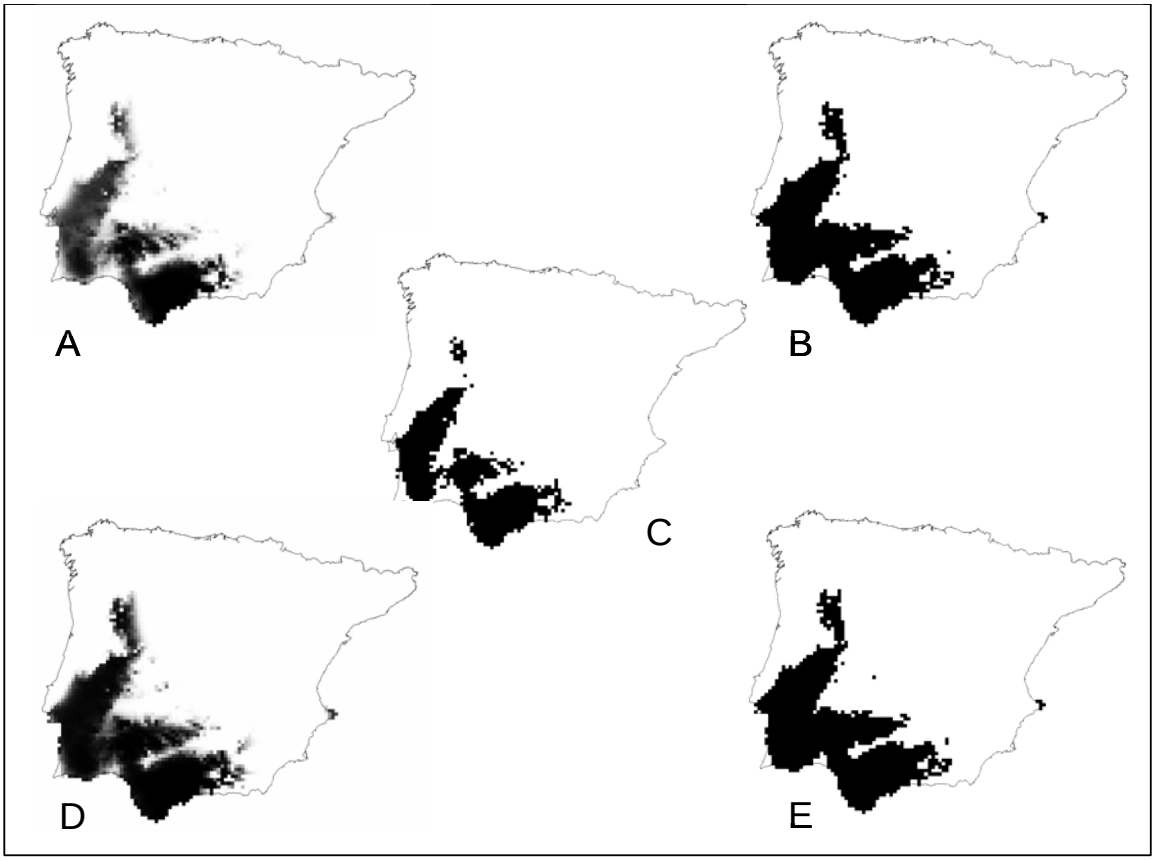



Figure 4.- A, logistic probabilities estimated for the Iberian Peninsula (dark grey indicate high values); B, presence-absence map after applying, to logistic probabilities, the cut-off that minimized the difference between sensitivity and specificity; C, presence-absence map after applying the 0.5 cut-off to the logistic probabilities; D, favourability values estimated after deleting the random element of the logistic equation (Real *et al*., in press); E, presence-absence map after applying the 0.5 cut-off to the favourability values.

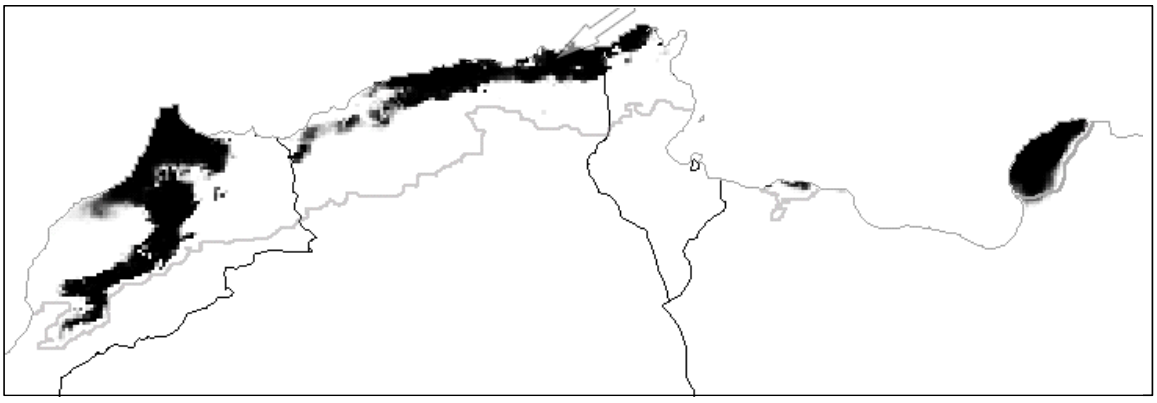


Figure 5.- Extrapolation of the environmental model of *Macrothele calpeiana* (Walckenaer, 1805) to North Africa. Values are favourability scores (dark grey indicate high values). Grey line delimits the area outside of which environmental variable are beyond the values used to parametrize the GLM model. The arrow points to El Arrouch, locality of Algeria from where Lucas (1849) cited *M. calpeiana*.

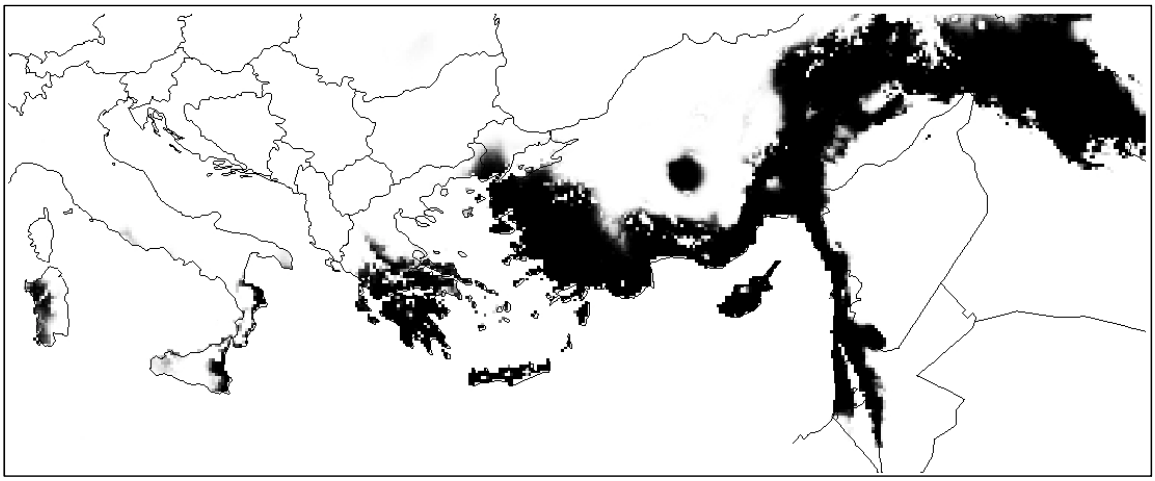


Figure 6.- Extrapolation of the environmental model of *Macrothele calpeiana* (Walckenaer, 1805) to the Mediterranean region. Values are favourability scores (dark grey indicate high values). The picture window delimits the area out of which environmental variables are beyond the values used to parametrize the GLM model.

Climate warming (Fig. 7) leads to a general reduction in potential Iberian Peninsula distribution area, from 11740 $km^2$ to 9600 $km^2$, affecting mainly in north of the Guadalquivir Valley, which nearly disappears as a distribution area, while potential distribution through the north of Portugal remains unaffected. There is a also a slight reduction of potential North African area, from 15520 $km^2$ to 14920 $km^2$, in spite of the potential Moroccan area increase (though fragmented in three) from 9040 $km^2$ to 9260 $km^2$.

## DISCUSSION

The climate envelope of a species can be considered the first constraint on its geographic distribution, at the top of scale-dependent, determinant factors (Mackey & Lindenmayer, 2001). Elucidation of that envelope is an essential first step in the understanding of any species distribution. For the moment, in the absence of detailed

physiological and ecological information, correlation methods such as the one used in the present study best further understanding of many species distributions. Although correlation does not prove cause, strength of relations and recurrent inter-species patterns seen can point to the agents that may be related with species distribution. In the case of *M. calpeiana*, precipitation seems to play a primary role in determining its geographic range. In particular, high seasonal precipitation variation seems to be the most relevant constraint; the greater the seasonal variation, the more favorable the habitat for the spider.

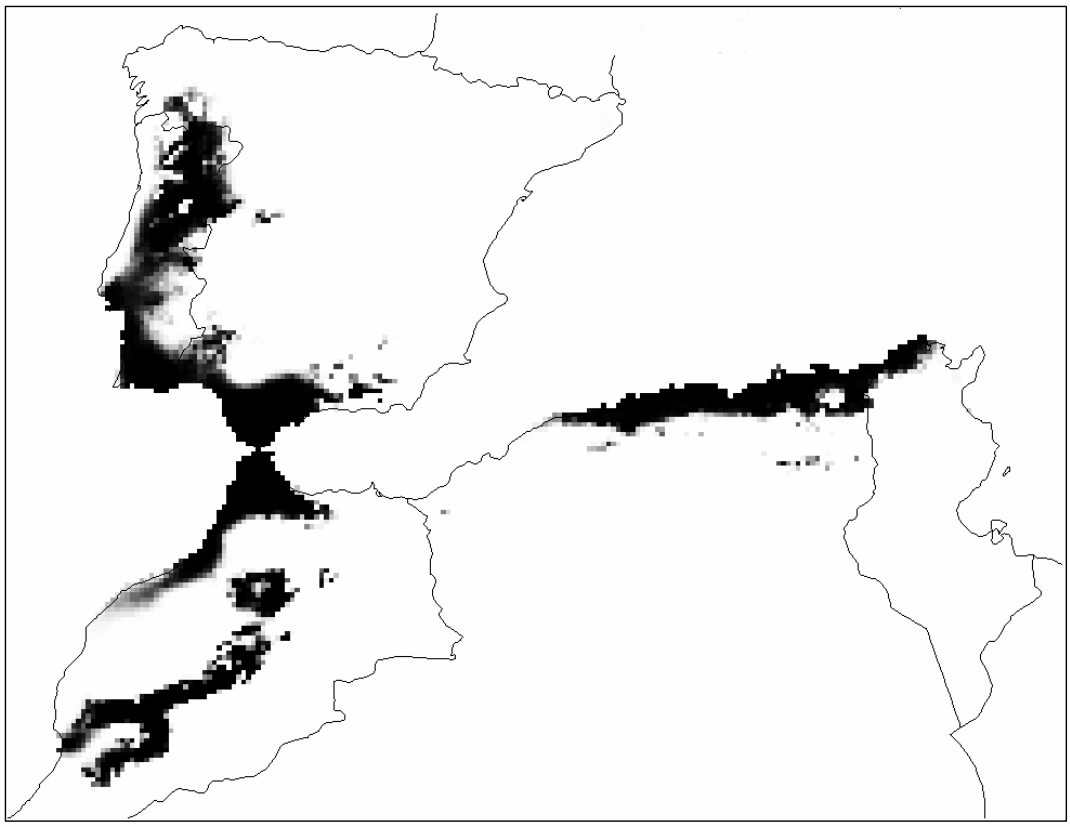


Figure 7.- Predicted changes in the potential distribution of *Macrothele calpeiana* (Walckenaer, 1805) in the Iberian Peninsula and North Africa after climate change. Values are favourability scores (dark grey indicate high values).

Helsdingen & Decae (1992) recognized the distribution gap in the Guadalquivir valley, and suggested that its open, unshaded areas could act as a barrier to *M. calpeiana*. In fact it is an ecological barrier, but primarily due to large-scale climate factors, as shown by our model. However, apart from the macroclimate, other factors could also have influenced; the south of the valley was under water during all the Tertiary and most of the Pleistocene (Martínez, 1989; Hevia, 2004), evidently a barrier to *M. calpeiana* colonization. Nevertheless, absence of *M. calpeiana* from this area has still to be confirmed.

The close fit of the potential model to the known distribution of the Iberian Peninsula cork oak, in the vitual absence of close spider-host plant association, indicates that the range of both species may be determined mainly by the same environmental factors, as suggested by Ferrández & Fernández de Céspedes (1996). Cork oak distribution, mainly in southern Atlantic Spain (Costa Tenorio *et al*., 1998), coincides with the two main populations of *M. calpeiana.* In Portugal, potential *M. calpeiana* distribution matches core areas of the cork oak, mainly in Alentejo province (see Fig. 29 in Costa Tenorio *et al*., 1998). The match of the distributions of other taxa with potential *M. calpeiana* distribution in the Iberian Peninsula suggests an environmental constraint common to all their distributions. This is the case, for example, of the leaf-beetle *Orestia punctipennis* (Lucas, 1849) (see Gruev & Döberl, 1997 and Baselga & Novoa, 2003), or the dung-beetle *Scarabaeus cicatricosus* (Lucas, 1846) (see Martín-Piera & López Colón, 2000; J. M. Lobo, unpublished data), among many others. These distribution areas coincide with the Iberian regions less affected during the last glacial period, which probably acted as a refuge for contracting species distribution (Carrión *et al.*, 2000).

Absence of records of *M. calpeiana* in Portugal is, at least, curious in the light of a possibly wider distribution of the spider in the past; particularly striking is the lack of observations of *M. calpeiana* in south and central Portugal, given its high degree of suitability and nearness to the second-most-important distribution area. It seems the Guadiana River has been a barrier to the dispersion of the spider in relatively recent times. Aerial dispersion (ballooning) in mygalomorph spiders is rare (Reichling, 2000); juveniles of *M. calpeiana* in particular are believed not to leave their nest far behind (Santos Lobatón, 1996). So, the barrier effect of the Guadiana River would seem plausible, given the apparently poor ability of the species to disperse. If its absence from Portugal is genuine, *M. calpeiana* is not in equilibrium with the environment, and its geographic range would be constrained by climate as well as by unique geographical and/or historical factors. However, if *M. calpeiana* distribution was wider in the past, as suggested by Ferrández & Fernández de Céspedes (1996), the species should be present, at least, in the well-conserved cork oak forests in south and central Portugal, where its extinction would seem unlikely. Confirmation of the absence of the spider from Portugal is urgently needed.

Two records, wrongly predicted by the model (see circle in Fig. 1), could indicate that these localities are at the species range limit; or alternatively, an important predictor variable could be lacking in our model, as autocorrelation in the residuals in the first lag distances seems to suggest. Failure by trend surface analysis to increase deviance explained by the environmental model could indicate that unaccounted-for variables may be acting on local scale (Diniz-Filho *et al*., 2003). Taxa poorly able to disperse, such as *M. calpeiana*, are highly influenced by local environmental conditions. Modelling such species data as spider presence at 100 $km^2$ resolution with larger-scale environmental predictors may fail to predict some records

that more precise geographical data and higher-resolution predictors could account for (Guisan & Hofer, 2003; Engler *et al*., 2004; Stockman *et al*., 2006). This does not mean that one scale of analysis is preferable to another, but that consideration of various scales is necessary for a more complete comprehension of the factors limiting species geographical distribution.

Although extrapolation of models from the area of their training data may be of interest, it can produce unreliable results; those from ostensibly direct variables, such as climate variables, can be more reliable than those from indirect ones (*sensu* Austin, 1980) such as altitude, usually a surrogate for other related factors whose relationships may vary according to region (Austin, 2002). Extrapolation from our model, based entirely on climate variables, presumably a direct influence on species physiology, should be more robust than extrapolation from others based on indirect variables. Nevertheless, truncated responses to some variables indicate incomplete characterization of those spider environmental requirements, so reliable extrapolation is limited to areas within training-area environmental ranges.

Extrapolation to North Africa identified environmentally suitable areas for *M. calpeiana* in that region, matching cork oak distribution there quite well, although forests of *Cedrus atlantica*, *Quercus pyrenaica* and *Pinus pinaster* are also common. That climate should be suitable in northern Algeria and Tunisia is quite interesting, as the Lucas (1849) record proceeded from within this potential territory (see arrow in Fig. 5). In 2003 and 2005, fourteen 1 $km^2$ localities in Morocco (four surveyors during ½ h., Fig. 8) were surveyed for *M. calpeiana*. The species was not found, not even in the four most northern localities, where highly suitable cork oak forests closely resemble ecosystems north of the Straits of Gibraltar. Given the relative ease of detection of the species (even though field absence reliabiltiy can never be complete),

these absences can be considered highly probably true. For the moment, given that presence of *M. calpeiana* in North Africa remains possible, but uncertian, and given the: i) potential of this area to sustain populations of the spider; ii) anthropogenic pressure lower there than in the Iberian Peninsula; and iii) uncertain taxonomy of central African *Macrothele* species; then all of the foregoing suggests a non-African origin for *M. calpeiana*. A probable East Antarctica origin for the Hexathelidae family (Raven, 1980), followed by radiation through eastern Gondwanaland and across Antarctica would explain the present Australian, Oriental, south and east Asian, and South American distribution of the family. Whereas an African origin for the *M. calpeiana* ancestor, followed by arrival of the spider on the Alboran plate, should not have led to the presumed lack of any species of the genus in North Africa. The most probable alternative hypothesis is that the Mediterranean was colonized from south Asia through a northern or southern Mediterranean dispersal route, during the Messinian salinity crisis (Sanmartín, 2003), perhaps during the late Oligocene-early Miocene (Oosterbroek & Arntzen, 1992; Rölg, 1999). As in the case of Portugal, North African presence (or absence) of *Macrothele* species must be clarified urgently, as the origin of the population from Ceuta must be, to elaborate a more reliable hypothesis on the origin and penetration route of this genus in the Iberian Peninsula.

Extrapolation to the Mediterranean area identified many suitable areas along coasts, mainly in the Greek and Anatolian mainland. Interestingly, Crete is identified as potentially suitable for *M. calpeiana*, while the other Mediterranean species of the same genus, *M. cretica*, is recorded from this island. In the Late Miocene, Crete belonged to a large land mass that the Tethys Sea later flooded, except for the highest altitudes (Welter-Schultes, 2000). Afterwards, in the Pliocene, uplifting of what had been many small islands gave Crete its present coastline. Given the favorability of the

area, survival of *M. cretica* or its predecessor on other Aegean Islands, or in the Greek and Anatolian Peninsulas, would not be surprising. Although the origin of *M. cretica* may be Asian (see above), on several occasions during the Miocene this species or its predecessor could have migrated from North Africa, perhaps from the Libyan region, towards the land masses that would become Crete (see Oosterbroek & Arntzen, 1992; Rölg, 1999). Thus, *M. cretica* or its predecessor could still survive in the potential areas identified in eastern North Africa. Another possibility is recent dispersal of *M. cretica* from North Africa or the Levant. Future phylogeographic studies will help to elucidate alternative regions of origin, and dispersal routes, of the genus.

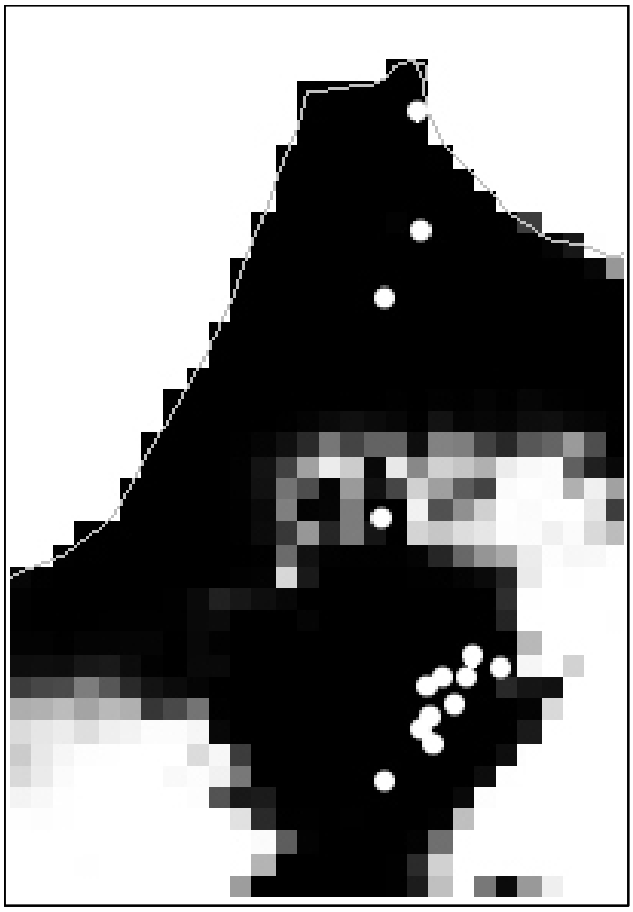


Figure 8.- North African 1 km$^2$ localities in which surveys were carried out in 2003 and 2005, without finding any specimen of *M. calpeiana*. The four most northern localities corresponded to cork oak forests, while the other localities were mostly *Cedrus atlanticus* forests.

Despite their limitations, model predictions are still the most reliable forecaster of climate change effects on species distributions (Pearson & Dawson, 2003; Martínez-Meyer, 2005). Our climate model should be considered the null hypothesis (Peterson *et al*., 2002) in the absence of other interactive factors such as, for example, biotic interactions, land use effects or historical factors. Potential habitat area for Iberian populations north of the Guadalquivir valley (core area 5; se Fig. 1) will be reduced, and that of core areas 2 and 3 will probably disappear under the climate change scenario considered. Although isolated populations could possibly persist in refuges undetected by our broad-resolution climate model (see Pearson, 2006), climate warming will not benefit such populations, in any case. Main core area 1 environment will remain highly favorable. In Portugal, potential habitat will be reduced in the south, while it will increase in the north. Given the known distribution of *M. calpeiana*, and its probably poor ability to disperse, it is not expected to undertake distribution shifts (along the Atlantic coast) to track optimal climatic conditions; its distribution area is predicted to be reduced, although the main core area will remain habitable. Thus, expansion of the potential distribution in north Portugal is practically irrelevant, unless new populations were discovered there in future. In North Africa, potential distribution will be reduced and fragmented, reducing potential habitat, especially in Morocco.

## FUTURE RESEARCH

Habitat modeling must be iterative, with new distribution data incorporated to the species database and models refitted until robust habitat patterns are obtained (Luck, 2002b). Given the favorability of North Africa for the species, of primary

importance is the corroboration of *M. calpeiana* presence or absence there, and also in Portugal. Survey designs such as those proposed by Jiménez-Valverde & Lobo (2004) and Hortal & Lobo (2005) should be used to maximize the environmental range sampled and, consequently, maximize field data usefulness for modeling. Model predictions, such as those developed in this study, can be used in the design process of the field survey (see Guisan *et al.*, 2006). True absence weight should be as great as presence weight in the modeling process, so sampling must be designed to provide reliable (though difficult to obtain) absences. Sampling designed to account for detectability (Mackenzie & Royle, 2005) can minimize the negative effect of false absences. Furthermore, detectability can be included as an explanatory factor in the model (Luck, 2002a; Royle *et al.*, 2005).

Model predictions can be used to support one biogeographic hypothesis or another, but statistics will never replace basic biological data. Thus, research priorities are to: i) confirm the presence or absence of the spider in Portugal and North Africa; ii) clarify the taxonomic status of central African *Macrothele* species; iii) estimate separation time among core Iberian populations, and between these Iberian populations and the population of Ceuta, via molecular data, and; iv) generate a reliable phylogenetic hypothesis for the genus. These are priorities for the formulation of a feasible, robust biogeographic hypothesis on the origin of *M. calpeiana*.

## AKCNOWLEDGEMENTS

This paper was supported by a MEC Project (CGL2004-04309), as well as by a Museo Nacional de Ciencias Naturales/C.S.I.C./Comunidad de Madrid PhD grant.

## LITERATURE CITED

Anderson, R. P., Gómez-Laverde, M. & Peterson, A. T. (2002) Geographical distributions of spiny pocket mice in South America: insights from predictive models. *Global Ecology and Biogeography*, **11**, 131-141.

Austin, M. P. (1980) Searching for a model for use in vegetation analysis. *Vegetatio*, **42**, 11-21.

Austin, M. P. (2002) Spatial prediction of species distribution: an interface between ecological theory and statistical modelling. *Ecological Modelling*, **157**, 101-118.

Barbosa, A. M., Real, R., Olivero, J. & Vargas, J. M. (2003) Otter (*Lutra lutra*) distribution modeling at two resolution scales suited to conservation planning in the Iberian Peninsula. *Biological Conservation*, **114**, 377-387.

Baselga, A. & Novoa, F. (2003) Los Chrysomelidae de los Arribes del Duero, noroeste de la Península Ibérica (Coleoptera). *Nouvelle Revue d'Entomologie (N. S.)*, **20**, 117-131.

Beaumont, L. J., Hughes, L. & Poulsen, M. (2005) Predicting species distributions: use of climatic parameters in BIOCLIM and its impact on predictions of species´ current and future distributions. *Ecological Modelling*, **186**, 250-269.

Blasco, A. & Ferrández, M. A. (1986) El género *Macrothele* Ausserer 1871 (Araneae; Dipluridae) en la Peninsula Ibérica. *Actas X Congreso Internacional de Aracnología Jaca/España*, **I**, 311-320.

Buckland, S. T., Burnham, K. P. & Augustin, N. H. (1997) Model Selection: An Integral Part of Inference. *Biometrics*, **53**, 603-618.

Busby, J. R. (1991) BIOCLIM – A Bioclimate Analysis and Prediction System. In *Nature Conservation: Cost Effective Biological Surveys and Data Analysis*, ed. C. R. Margules & M. P. Austin, pp. 64-68. CSIRO, Australia.

Bustamante, J. & Seoane, J. (2004). Predicting the distribution of four species of raptors (Aves: Accipitridae) in southern Spain: statistical models work better than existing maps. *Journal of Biogeography*, **31**, 295-306.

Carrión, J. S., Parra, I., Navarro, C. & Munuera, M. (2000) Past distribution and ecology of the cork oak (*Quercus suber*) in the Iberian Peninsula: a pollen-analytical approach. *Diversity and Distributions*, **6**, 29-44.

Costa Tenorio, M., Morla Juaristi, C. & Sainz Ollero, H. (eds.) (1998). *Los bosques ibéricos. Una interpretación geobotánica.* Geoplaneta, Barcelona.

Cramer, J. S. (1999) Predictive performance of binary logit model in unbalanced samples. *Journal of the Royal Statistical Society: Series D (The Statistican)*, **48**, 85-94.

Diniz-Filho, J. A. F., Bini, L. M., Hawkins, B. A. (2003) Spatial autocorrelation and red herrings in geographical ecology. *Global Ecology & Biogeography*, **12**, 53-64.

Elith, J., Graham, C. H., Anderson, R. P., Dudík, M., Ferrier, S., Guisan, A., Hijmans, R. J., Huettmann, F., Leathwick, J. R., Lehmann, A., Li, J., Lohmann, L.G., Loiselle, B. A., Manion, G., Moritz, C., Nakamura, M., Nakazawa, Y., Overton, J. M., Peterson, A. T., Phillips, S. J., Richardson, K., Scachetti-Pereira, R., Schapire, R. E., Soberón, J., Williams, S., Wisz, M. S. and Zimmermann, N. E. (2006) Novel methods improve prediction of species' distributions from occurrence data. *Ecography*, **29**, 129-151.

Engler, R., Guisan, A. & Rechsteiner, L. (2004) An improved approach for predicting the distribution of rare and endangered species from occurrence and pseudo-absence data. *Journal of Applied Ecology*, **41**, 263-274.

Feinstein, A. R. (1996) *Multivariable analysis: an introduction*. Yale University Press, New Haven.

Ferrández, M. A. & Fernández de Céspedes, H. 1996. *Macrothele calpeiana* (Walckenaer, 1805). In *Los Invertebrados no insectos de la "Directiva Hábitat" en España*, eds. M. A. Ramos, D. Bragado & J. Fernández, pp. 129-141. Dirección General de Conservación de la Naturaleza.

Ferrier, S., & Watson, G. (1997) *An evaluation of the effectiveness of environmental surrogates and modelling techniques in predicting the distribution of biological diversity.* Environment Australia, Canberra, available in http://www.deh.gov.au/biodiversity/publications/technical/surrogates/

Gallego, D., Cánovas, F., Esteve, M. A. & Galián, J. (2004) Descriptive biogeography of *Tomicus* (Coleoptera: Scolytidae) species in Spain. *Journal of Biogeography*, **31**, 2011-2024.

Govindasamy, B., Duffy, P. B. & Coquard, J. (2003) High-resolution simulations of global climate, part 2: effects of increased greenhouse cases. *Climate Dynamics*, **21**, 391-404.

Gruev, B. & Döberl, M. (1997) General distribution of the flea beetles in Palaearctic subregion (Coleoptera, Chrysomelidae: Alticinae). *Scopolia*, **37**, 1-496.

Guisan, A. & Zimmermann, N. E. (2000) Predictive habitat distribution models in ecology. *Ecological Modelling*, **135**, 147-186.

Guisan, A., Edwards, T. C. & Hastie, T. (2002) Generalized linear and generalized additive models in studies of species distributions: setting the scene. *Ecological Modelling*, **157**, 89-100.

Guisan, A. & Hofer, U. (2003) Predicting reptile distributions at the mesoscale: relation to climate and topography. *Journal of Biogeography*, **30**, 1233-1243.

Guisan, A. & Thuiller, W. (2005) Predicting species distributions: offering more than simple habitat models. *Ecology Letters*, **8**, 993-1009.

Guisan, A., Broennimann, O., Engler, R., Yoccoz, N. G., Vust, M., Zimmermann, N. E., Lehmann, A. (2006) Using niche-based models to improve the sampling of rare species. *Conservation Biology*, **20**, 501-511.

Harrell, F. E. J. 2001. *Regression modelling strategies: with applications to linear models, logistic regression, and survival analysis*. Springer, NY.

Hastie, T. J. & Tibshirani, R. J. (1990) *Generalized Additive Models*. Chapman & Hall, London.

Helsdinge, P. J. van & Decae, A. E. (1992) Ecology, distribution and vulnerability of *Macrothele calpeiana* (Walckenaer) (Araneae, Hexathelidae). *Tijdschrift voor Entomologie*, **135**, 169-178.

Hevia, I. M. (2004) *Geología de España. Una historia de seiscientos millones de años*. Rueda S. L., Madrid.

Hickling, R., Roy, D. B., Hill, J. K., Fox, R. & Thomas, C. D. (2006) The distributions of a wide range of taxonomic groups are expanding polewards. *Global Change Biology*, **12**, 450-455.

Hijmans, R. J., Cameron, S. E., Parra, J. L., Jones, P. G. & Jarvis, A. (2005) Very high resolution interpolated climate surfaces for global land areas. *International Journal of Climatology*, **25**, 1965-1978.

Hirzel, A. H., Posse, B., Oggier, P.-A., Crettenand, Y., Glenz, C., Arlettaz, R. (2004) Ecological requirements of a reintroduced species, with implications for release policy: the Bearded vulture recolonizing the Alps. *Journal of Applied Ecology*, **41**, 1103-1116.

Hortal, J. & Lobo, J. M. (2005) An ED-based protocol for optimal sampling of biodiversity. *Biodiversity and Conservation*, **14**, 2913-2947.

Hulme, P. E. (2005) Adapting to climate change: is there scope for ecological management in the face of a global threat? *Journal of Applied Ecology*, **42**, 784-794.

IPCC (2001) *Climate change 2001: the scientific basis*. Cambridge University Press, Cambridge

Jiménez-Valverde, A. & Lobo, J. M. (2004) Un método sencillo para seleccionar puntos de muestreo con el objeto de inventariar taxones hiperdiversos: el caso práctico de las familias *Araneidae* y *Thomisidae* (*Araneae*) en la Comunidad de Madrid, España. *Ecología*, **18**, 297-308.

Jiménez-Valverde, A. & Lobo, J. M. (2006) The ghost of unbalanced species distribution data in geographic model predictions. *Diversity and Distributions*, in press.

Jiménez-Valverde, A., Ortuño, V. M. & Lobo, J. M. Exploring the distribution of *Sterocorax* Ortuño, 1990 (Coleoptera, Carabidae) species in the Iberian Peninsula. *Journal of Biogeography*, in press.

King, D. (2005) Climate change: the science and the policy. *Journal of Applied Ecology*, **42**, 779-783.

Legendre, P. & Legendre, L. (1998) *Numerical Ecology.* Elsevier, Amsterdam.

Lehmann, A., Overton, J. McC. & Austin, M. P. (2002a) Regression models for spatial prediction: their role for biodiversity and conservation. *Biodiversity and Conservation*, **11**, 2085-2092.

Lehmann, A., Overton, J. McC. & Leathwick, J. R. (2002b) GRASP: generalized regression analysis and spatial prediction. *Ecological Modelling*, **157**, 189-207.

Lobo, J. M., Verdú, J. R. & Numa, C. (2006). Environmental and geographical factors affecting the Iberian distribution of flightless *Jekelius* species (Coleoptera: Geotrupidae). *Diversity and Distributions*, **12**, 179-188.

Lovejoy, T. E. & Hannah, L. (eds.) (2005) *Climate change and biodiversity*. Yale University Press, New Haven & London.

Lucas, H. (1849) *Exploration scientifique de l'Algérie pendant les années 1840, 1841, 1842, publiée par ordre du gouverments. Sciences physiques. Zoologie. Histoire naturelle des animaux articulés.* Paris.

Luck, G. W. (2002a) The habitat requirements of the rufous treecreeper (*Climacteris rufa*). 1. Preferential habitat use demonstrated at multiple spatial scales. *Biological Conservation*, **105**, 383-394.

Luck, G. W. (2002b) The habitat requirements of the rufous treecreeper (*Climacteris rufa*). 2. Validating predictive habitat models. *Biological Conservation*, **105**, 395-403.

Mackenzie, D. I. & Royle, J. A. (2005) Designing occupancy studies: general advice and allocating survey effort. *Journal of Applied Ecology*, **42**, 1105-1114.

Mackey, B. G. & Lindenmayer, D. B. (2001) Towards a hierarchical framework for modelling the spatial distribution of animals. *Journal of Biogeography*, **28**, 1147-1166.

Malcom, J. R., Liu, C., Neilson, R. P., Hansen, L. & Hannah, L. (2006) Global warming and extinctions of endemic species from biodiversity hotspots. *Conservation Biology*, **20**, 538-548.

Manel, S., Días, J. M. & Ormerod, S. J. (1999a) Comparing discriminant analysis, neural networks and logistic regression for predicting species distributions: a case study with a Himalayan river bird. *Ecological Modelling*, **120**, 337-347.

Manel, S., Dias, J. M., Buckton, S. T. & Ormerod, S. J. (1999b) Alternative methods for predicting species distributions: an illustration with Himalayan river birds. *Journal of Applied Ecology*, **36**, 734-747.

Martín-Piera, F. & López-Colón, J. I. (2000) *Coleoptera, Scarabaeoidea I* (ed. M. A. Ramos), Fauna Ibérica, vol. 14. Museo Nacional de Ciencias Naturales, Consejo Superior de Investigaciones Científicas, Madrid.

Martínez, N. L. (1989) Tendencias en Paleobiogeografía. In *Paleontología, nuevas tendencias*, ed. E. Aguirre, pp. 271-296. Consejo Superior de Investigaciones Científicas, Madrid.

Martínez-Meyer, E. (2005) Climate change and biodiversity: some considerations in forecasting shifts in species´ potential distributions. *Biodiversity Informatics*, **2**, 42-55.

McCullagh, P. & Nelder, J. A. (1997) *Generalized Linear Models.* Monographs on Statistics and Applied Probability. Chapman & Hall, London.

Moisen, G. G. & Frescino, T. S. (2002) Comparing five modelling techniques for predicting forest characteristics. *Ecological Modelling*, **157**, 209-225.

Muñoz, A. R., Real, R., Barbosa, A. M. & Vargas, J. M. (2005) Modelling the distribution of Bonelli´s eagle in Spain: implications for conservation planning. *Diversity and Distributions*, **11**, 477-486.

Olden, J. D., Jackson, D. A. & Peres-Neto, P. (2002) Predictive models of fish species distributions: a note on proper validation and chance predictions. *Transactions of the American Fisheries Society*, **131**, 329-336.

Olivier, F. & Wotherspoon, S. J. (2005) GIS-based application of resource selection functions to the prediction of snow petrel distribution and abundance in East Antarctica: Comparing models at multiple scales. *Ecological Modelling*, **189**, 105-129.

Oosterbroek, P. & Arntzen, J. W. (1992) Area-Cladograms of Circum-Mediterranean taxa in relation to Mediterranean palaeogeography. *Journal of Biogeography*, **19**, 3-20.

Oreskes, N. (2004) The scientific consensus on climate change. *Science*, **306**, 1686.

Parmesan, C. & Yohe, G. (2003) A globally coherent fingerprint of climate change impacts across natural systems. *Nature*, **421**, 37-42.

Pearson, R. G. (2006) Climate change and the migration capacity of species. *Trends in Ecology and Evolution*, **21**, 111-113.

Pearson, R. G. & Dawson, T. P. (2003) Predicting the impacts of climate change on the distribution of species: are bioclimate envelope models useful? *Global Ecology and Biogeography*, **12**, 361-371.

Peterson, A. T. (2003) Projected climate change effects on Rocky Mountain and Great Plains birds: generalities of biodiversity consequences. *Global Change Biology*, **9**, 647-655.

Peterson, A. T., Ortega-Huerta, M. A., Bartley, J., Sánchez-Cordero, V., Soberón, J., Buddemeier, R. H. & Stockwell, D. R. B. (2002) Future projections for Mexican faunas under global climate change scenarios. *Nature*, **416**, 626-629.

Platnick, N. (2006) *The World Spider Catalogue v 6.5.* American Museum of Natural History, in http://research.amnh.org/entomology/spiders/catalog/INTRO1.html

R Development Core Team (2004) *R: A language and environment for statistical computing.* R Foundation for Statistical Computing, Vienna, Austria. Available at http://www.R-project.org

Raven, R. J. (1980) The evolution and biogeography of the mygalomorph spider family Hexathelidae (Araneae, Chelicerata). *Journal of Arachnology*, **8**, 251-266.

Raxworthy, C. J., Martínez-Meyer, E., Horning, N., Nussbaum, R. A., Schbeider, G. E., Ortega-Huerta, M. A., Peterson, A. T. (2003) Predicting distributions of known reptile species in Madagascar. *Nature*, **426**, 837-841.

Real, R., Barbosa, A. M. & Vargas, J. M. Obtaining environmental favourability functions from logistic regression. *Environmental and Ecological Statistics*, in press.

Reichling, S. B. (2000) Group dispersal in juvenile *Brachypelma vegans* (Araneae, Theraphosidae)**.** *Journal of Arachnology*, **28**, 248-250.

Reineking, B. & Schröder, B. (2003) Computer-intensive methods in the analysis of species-habitat relationships. In *GfÖ Arbeitskreis Theorie in der Ökologie: Gene, Bits und Ökosysteme*, ed. H. Reuter, B. Breckling & A. Mittwollen, pp. 100-117. P. Lang Verlag Frankfurt/M.

Rölg, F. (1999) Mediterranean and Parathethys. Facts and hypothesis of an Oligocene to Miocene paleogeography (short overview). *Geologica Carpathica*, **50**, 339-349.

Root, T. L., Price, J. T., Hall, K. R., Schneider, S. H., Rosenzweig, C. & Pounds, J. A. (2003) Fingerprints of global warming on wild animals and plants. *Nature*, **421**, 57-60.

Royle, J. A., Nichols, J. D. & Kéry, M. (2005) Modelling occurrence and abundance of species when detection is imperfect. *Oikos*, **110**, 353-359.

Russell, K. R., Mabee, T. J. & Cole, M. B. (2004) Distribution and habitat of Columbia Torrent Salamanders at multiple spatial scales in managed forests of Northwestern Oregon. *Journal of Wildlife Management*, **68**, 403-415.

Sanmartín, I. (2003) Dispersal vs. vicariance in the Mediterranean: historical biogeography of the Paleartic Pachydeminae (Coleoptera, Scarabaeoidea). *Journal of Biogeography*, **30**, 1883-1897.

Santos Lobatón, M. C. (1996) Estudio sobre *Macrothele calpeiana* Walckenaer, 1805 (Araneae, Hexathelidae) en dos pinares de la provincia de Cádiz (España). *Aracnología*, **24**, 1-10.

Schadt, S., Revilla, E., Wiegand, T., Knauer, F., Kaczensky, P., Breitenmoser, U., Bufka, L., Červený, J., Koubek, P., Huber, T., Staniša, C. & Trepl, L. (2002) Assessing the suitability of central European landscapes for the reintroduction of Eurasian lynx. *Journal of Applied Ecology*, **39**, 189-203.

Scott, J. M., Heglund, P. J., Haufler, J. B., Morrison, M., Raphael, M. G., Wall, W. B. & Samson, F. (eds.) (2002) *Predicting Species Occurrences. Issues of Accuracy and Scale*). Island Press, Covelo, CA.

Seoane, J., Bustamante, J. & Díaz-Delgado, R. (2005) Effects of expert opinion on the predictive ability of environmental models of bird distribution. *Conservation Biology*, **19**, 512-522.

Silva, L. C. & Barroso, I. M. (2004) *Regresión logística*. La Muralla, Madrid.

Snazell, R. (1986) The spider genus *Macrothele* Ausserer in Spain (Araneae; Dipluridae). *Bulletin of the British Ecological Society*, **17**, 80-83.

Snazell, R. & Allison, R. (1989) The genus *Macrothele* Ausserer (Araneae, Hexathelidae) in Europe. *Bulletin of the British Arachnological Society*, **8(3)**, 65-72.

Sawada, M. (1999) ROOKCASE: an Excel 97/2000 Visual Basic (VB) Add-in for exploring global and local spatial autocorrelation. *Bulletin of the Ecological Society of America*, **80**, 231–234.

Stockman, A. K., Beamer, D. A. & Bond, J. E. (2006) An evaluation of a GARP model as an approach to predicting the spatial distribution of non-vagile invertebrate species. *Diversity and Distributions*, **12**, 81-89.

Thomas, C. D., Cameron, A., Green, R. E., Bakkenes, M., Beaumont, L. J., Collingham, Y. C., Erasmus, B. F. N., Ferreira de Siqueira, M., Grainger, A., Hannah, L., Hughes, L., Huntley, B., van Jaarsveld, A. S., Midgley, G. F., Miles, L., Ortega-Huerta, M. A., Peterson, A. T., Phillips, O. L. & Williams, S. E. (2004) Extinction risk from climate change. *Nature*, **427**, 145-148.

Thuiller, W., Lavorel, S., Araújo, M. B., Sykes, M. T. & Prentice, I. C. (2005a) Climate change threats to plant diversity in Europe. *Proceedings of the National Academy of Sciences, USA*, **102**, 8245-8250.

Thuiller, W., Lavorel, S. & Araújo, M. B. (2005b) Niche properties and geographical extent as predictors of species sensitivity to climate change. *Global Ecology and Biogeography*, **14**, 347-357.

Welter-Schultes, F. W. (2000) The paleogeography of late Neogene central Crete inferred from the sedimentary record combined with *Albinaria* land snail biogeography. *Palaeogeography, Palaeoclimatology, Palaeoecology*, **157**, 27-44.

Wintle, B. A., Elith, J. & Potts, J. M. (2005) Fauna habitat modelling and mapping: a review and case study in the Lower Hunter Central Coast region of NSW. *Austral Ecology*, **30**, 719-738.

Wood, S.N. (2000) Modelling and smoothing parameter estimation with Multiple Quadratic Penalties. *Journal of the Royal Statistical Society Series B*, **62(2)**, 413-428.

Wood, S. N. (2004) *mgcv: GAMs with GCV smoothness estimation and GAMMs by REML/PQL*. R package version 1.1-8.

Wood, S. N. & Augustin, N. H. (2002) GAMs with integrated model selection using penalized regression splines and applications to environmental modelling. *Ecological Modelling*, **157**, 157-177.

Zaniewski, A. E., Lehmann, A. & Overton, J. McC. (2002) Predicting species spatial distributions using presence-only data: a case study of native New Zealand ferns. *Ecological Modelling*, **157**, 261-280.

**Annex. —** Distribution data sources of *Macrothele calpeiana* in the Iberian Peninsula.

Blasco, A. & Ferrández, M. A. (1986) El género *Macrothele* Ausserer 1871 (Araneae; Dipluridae) en la Peninsula Ibérica. *Actas X Congr. Int. Aracnol. Jaca/España*, **I**, 311-320.

Calvo-Hernández, D. & Santos-Lobatón, M. C. (2001) Variabilidad morfológica en las poblaciones de *Macrothele calpeiana* (Walckenaer, 1805) (Araneae, Hexathelidae) en la provincia de Cádiz (España). *Revista Ibérica de Aracnología*, **3**, 43-45.

Calzada, J. (2002) Presencia de *Macrothele calpeiana* (Walckenaer, 1805) en las inmediaciones de la Reserva Natural del Peñon de Zaframagón (España). *Revista Ibérica de Aracnología*, **5**, 83-84.

Fernández, M. A. (2004) *Macrothele calpeiana* (Walckenaer, 1805). Situación actual y perspectivas. *Munibe (Suplemento)*, 21, 154-161. [Map 1 in p. 157].

Helsdinge, P. J. van & Decae, A. E. (1992) Ecology, distribution and vulnerability of *Macrothele calpeiana* (Walckenaer) (Araneae, Hexathelidae). *Tijdschrift voor Entomologie*, **135**, 169-178.

Luque, F. J. R. (2001) Nuevos datos de *Macrothele calpeiana* (Walckenaer, 1805) para Jaén (España). *Revista Ibérica de Aracnología*, **4**, 34.

Rodríguez, E. D. & García-Villanueva, V. (2000) Primeros datos sobre la presencia de *Macrothele calpeiana* (Walckenaer, 1805) en Extremadura (España). *Revista Ibérica de Aracnología*, **1**, 57-58.

Santos Lobatón, M. C. (1996) Estudio sobre *Macrothele calpeiana* Walckenaer, 1805 (Araneae, Hexathelidae) en dos pinares de la provincia de Cádiz (España). *Aracnología*, **24**, 1-10.

Snazell, R. & Allison, R. (1989) The genus *Macrothele* Ausserer (Araneae, Hexathelidae) in Europe. *Bulletin of the British Arachnological Society*, **8(3)**, 65-72.

# FACTORES DETERMINATES DE LA DISTRIBUCIÓN DEL ENDEMISMO IBÉRICO *MACROTHELE CALPEIANA* (WALCKENAER, 1805) (ARANEAE, HEXATHELIDAE)

**RESUMEN.** Se conoce poco sobre los requerimientos de hábitat de *Macrothele calpeiana* (Walckenaer, 1805), araña endémica de la Península Ibérica. Este trabajo pretende identificar los posibles determinantes de su distribución, tratando de separar sus efectos puros de los combinados. Los datos disponibles de presencia de la especie se modelizaron usando Modelos Generalizados Lineales (GLMs) y un conjunto de variables relacionadas con el clima, el uso de suelo y el vigor de la vegetación. Los efectos puros y combinados se estimaron mediante partición de la varianza y partición jerarquizada. A la escala de este estudio, la distribución de *M. calpeiana* está principalmente determinada por variables climáticas, especialmente por aquellas relacionadas con la precipitación. La especie se ve favorecida por precipitaciones anuales elevadas y por un alto grado de estacionalidad. La temperatura también es importante, siendo evitados los valores extremos. A pesar de que los efectos puros del vigor vegetal y del uso del suelo son insignificantes, la pérdida de masas forestales en favor de tierras agrícolas parece tener un efecto negativo sobre la araña. Se discute la incapacidad del modelo climático para predecir algunas áreas de distribución de la especie. Se resalta la necesidad de disponer de datos de distribución de alta calidad para desarrollar modelos de distribución fiables.

**Palabras clave:** clima, modelos de hábitat, Península Ibérica, uso del suelo, *Macrothele calpeiana*

**Este capítulo ha sido enviado a publicar como:**

JIMÉNEZ-VALVERDE, A. & LOBO, J. M. Distribution determinants of endangered Iberian spider *Macrothele calpeiana* (Araneae, Hexathelidae). *Environmental Entomology.*

# DISTRIBUTION DETERMINANTS OF ENDANGERED IBERIAN SPIDER *MACROTHELE CALPEIANA* (ARANEAE, HEXATHELIDAE)

**ABSTRACT.** Little is known about the habitat preferences of *Macrothele calpeiana* (Walckenaer, 1805), an endangered endemic Iberian spider. This work seeks to identify its possible distribution determinants, trying to disentangle their independent from combined effects. Generalized Linear Models (GLMs) of species presence-absence in southern Iberia were built from available distribution information and a variety of climate, land-use and vegetation-vigor explanatory variables. Their independent and combined effects were estimated using variation and hierachical partitioning. On the scale of this work, *M. calpeiana* distribution is determined mainly by climate variables, especially by those related with precipitation; high annual precipitation and high precipitation periodicity favours the spider. Temperature is also important, as the species is not found where temperatures reach extremes. While independent vegetative vigor and land-use effects, not easily separated from climate effects, are negligible, loss of forest to agriculture seems to have a negative effect. Failure of climate model interpolation to predict some core species distribution areas in southern Iberia is discussed. The need for reliable distribution information from which to develop accurate habitat models is highlighted.

**Keywords:** climate, habitat models, Iberian Peninsula, land use, *Macrothele calpeiana*

## INTRODUCTION

*Macrothele calpeiana* (Walckenaer, 1805), an endemic Iberian spider included in the Bern and Habitat directives, is distributed solely in southern Spain. Practically all populations have been found in the Guadalquivir river basin (see Fig 1), with the exception of a North African (Ceuta) record, considered to be the result of recent introduction by Spain-Morocco maritime traffic (Ferrández & Fernández de Céspedes, 1996).

*M. calpeiana*, a non-vagile, long-lived spider (females can live longer than 5 years; Perry, 2002), spins an aerial sheet-web which continues in underground silk tubes (funnel–webs), usually under stones or roots, but also in holes and crevices in bare ground, and even under tree bark several meters above ground (Gallon, 1994; Santos Lobatón, 1996). Mating seems to occur mainly in spring (May-March); spiderlings emerge in summer (August), probably remaining in the maternal retreat until October (Snazell & Allison, 1989; Perry, 2002).

*M. calpeiana* populations, found mainly in cork oak (*Quercus suber*) forests (Snazell, 1986; Snazell & Allison, 1989; Helsdinge & Decae, 1992), where winters are warm, summer temperatures high and rainfall copious, find a variety of habitats suitable (e.g., scrub land, pine forests, eucalyptus plantations; Helsdinge & Decae, 1992). Supplying some non-systematic density information, Helsdinge & Decae, 1992, speculated that *M. calpeiana* was favoured by a moderate amount of anthropogenic activity, and did not consider the species to be an indicator of cork oak forests. However, studies have not been designed to provide conclusive information on *M. calpeiana* habitat preferences, so that its habitat requirements remain practically unknown.

This paper explores, at a resolution of 1 × 1 km, the determinants of *M. calpeiana* distribution in the southern Iberian Peninsula, its main distribution area. Taking the place of experimentation, impossible on large spatial scales, habitat modelling (Guisan & Zimmermann, 2000; Guisan & Thuiller, 2005), the technique used in this study, tests most effectively for species habitat preferences, while identifying major predictor variables most reliably. As explanatory variable correlation is an obstacle to the determination of probable causal factors, effects of climate, land-use and vegetation-vigor groups were investigated via variation partitioning (Legendre & Legendre, 1998); independent effects of single variables were investigated via hierarchical partitioning (Chevan & Sutherland, 1991).

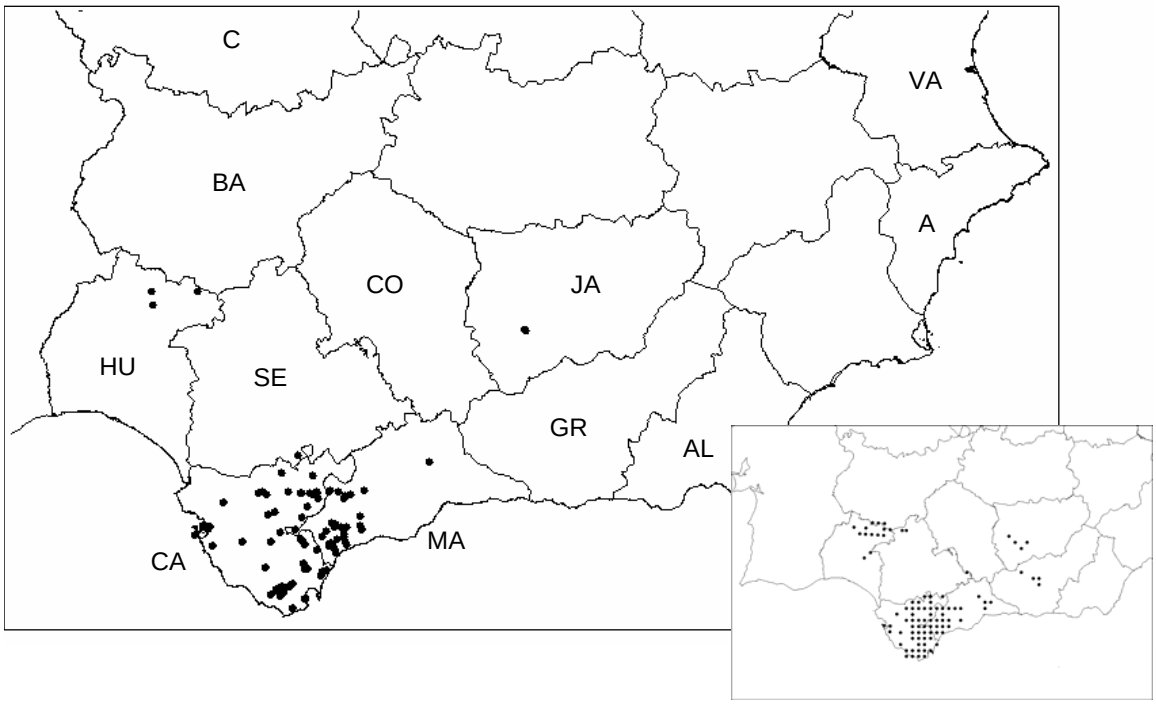



Figure 1.- Records of *Macrothele calpeiana* (Walckenaer, 1805) in the Iberian Peninsula referred to 1×1 km UTM squares. Spanish provinces cited in the text: A, Alicante; AL, Almería; BA, Badajoz; C, Cáceres; CA, Cádiz; CO, Córdoba; GR, Granada; HU, Huelva; JA, Jaén; MA, Málaga; SE, Sevilla; VA, Valencia. Window showing the 10×10 km presence points available (92); some core areas are not represented in the 1×1 km presence data.

## METHODS

**Biological data and extent of study. —** *M. calpeiana* presence data in 1 $km^2$ UTM squares were extracted from the literature (see Annex); 89 presence points were available for the species in the southern Iberian Peninsula (maximum latitude 39º 39'; see Fig. 1). It is necessary to note that some populations only had distribution data at a resolution of 10 $km^2$ and no 1 $km^2$ UTM square was available (see window in Fig. 1).

**Environmental data. —** At a resolution of 1×1 km in southern Spain, eight climate variables were considered: yearly days of frost; insolation (annual hours of sunlight); annual precipitation, precipitation periodicity (the coefficient of variation of monthly scores); mean annual temperature; minimum winter temperature; maximum summer temperature and annual temperature range. All these variables were courtesy of the Spanish Instituto Nacional de Meteorología, while mean altitude, slope and aspect were obtained from a Digital Elevation Model, at a resolution of 3 arc-seconds (~90 meters), provided by the United States geological Survey (http://www.usgs.gov). The mean Normalized Difference Vegetation Index (NDVI, the photosynthetically active radiation that plants absorb, is a measure of plant density and vegetation health; Chong *et al*., 1993; Pettorelli *et al.*, 2005) for year 2001 was provided by the Instituto Nacional de Técnica Aeroespacial (CREPAD, Gran Canaria, Spain); to minimize cloud and/or aerosol noise, mean annual NDVI was calculated from maximum monthly values. For each 1×1 km UTM square, the percentage of: forest; agricultural land; scrub and grassland; open places with little or no vegetation; and artificial surfaces were extracted from the Corine Land Cover 2000 (100 m resolution; see http://terrestrial.eionet.europa.eu/CLC2000)

**Statistical analyses. —** Tabulated maximum and minimum values of the 17 above-mentioned variables at presence points defined the multi-dimensional envelope for *M. calpeiana* (see Busby, 1986 and Lobo *et al*., 2006). From the area outside the envelope, 801 pseudo-absences were randomly selected (prevalence=0.1). As absences from a biological atlas are not necessarily true absences (the species may be present in a particular cell but not recorded), their inclusion as erroneous data reduces model prediction power; their substitution by pseudo-absences limits the amount of noise in the data (Jiménez-Valverde *et al*., in press). Moreover, these pseudo-absences may be used with prediction techniques employing both presence and absence data, to enhance prediction accuracy (Zaniewski *et al*., 2002; Engler *et al*., 2004; Lobo *et al*., 2006).

Generalized Additive Models (GAMs) with penalized regression splines (Wood, 2000; Wood & Augustin, 2002) were used to explore spider presence-absence relationships with predictors. The mgcv package (Wood, 2004) fitted GAMs, with four initial degrees of freedom, in R (R Development Core Team, 2004). To reduce the effects of multi-collinearity, predictors were first classified in intra-correlated groups by means of an $r \geq 0.8$ classifier threshold ($r$, Pearson´s correlation coefficient; Silva & Barroso, 2004). The members of the group explaining less GAM deviance or with a complex or unrealistic relationship with presence-absence data were then dropped.

Occurrence of *M. calpeiana* was finally modeled using logistic regression analysis (Generalized Linear Models with binomial distribution and logit-link function; McCullagh & Nelder, 1997). Models were backward-stepwise fitted (Harrell, 2001), producing nested models to be AIC-tested (Buckland *et al*., 1997), a penalization of the log-likelihood of the model as function of the number of degrees

of freedom. GLMs were fitted in R (R Development Core Team, 2004). Probabilities produced by logistic regression, unavoidably biased towards the most common event (Cramer, 1999), were corrected with a favourability function (Real *et al*., in press): the favourability probabilities so derived were then mapped. Residuals of the logistic functions were examined and tested for autocorrelation using the Moran´s *I* spatial autocorrelation statistic (Sawada, 1999), selecting a lag distance of 12 kilometers. Moran's *I* test was checked for significance with the Bonferroni-corrected significance level. Spatial autocorrelation in the residuals usually indicates that the model must be enlarged to incorporate spatially structured variables not otherwise accounted for (Odland, 1988); addition of complex spatial terms (the third-degree polynomial of latitude and longitude) to the model can be expected to account for those ignored variables.

Models were "leave-out-one", jackknife validated; i.e., one observation was excluded, the model parametrized again with the remaining *n*-1 observations, a predicted probability obtained for the excluded observation, and the procedure repeated *n* times (see Olden *et al*., 2002). With these new jackknife probabilities, the area under the ROC curve (AUC), a measure of overall discriminatory power (Fielding & Bell, 1997), was computed. Also, sensitivity (presences correctly predicted) and specificity (absences correctly predicted) were calculated using the threshold which minimizes their difference (Jiménez-Valverde & Lobo, 2006). All validation computations were run in R (R Development Core Team, 2004).

Although previously corrected for multi-collinearity, explanatory variables remain unavoidably correlated. In the assessment of the relative influence of each group of climate, land-use and NDVI explanatory variables on *M. calpeiana* presence, variation partitioning (Legendre & Legendre, 1998) determined the independent

effects of: (a) climate variation alone; (b) land use alone; (c) NDVI alone; and the combined effects of; (d) climate and land-use components; (e) climate and NDVI components; (f) land use and NDVI components; (g) the three components. Deviance reduction

$$D = [\text{Null deviance-Residual deviance}]/\text{Null deviance}$$

can be used to compare models from different combinations of factors (Guisan & Zimmermann, 2000). Total deviance (*D*) is obtained by regressing the dependent variable against the three groups of factors. Percentage of explained deviance is also computed for pairs of variables and for each variable alone. The independent effect of each group of variables is obtained by subtracting variation explained by the combination of the other groups, from variation explained by the combination of all explanatory variables together. Variation attributable to the combined effect of pairs of groups may be obtained by simple sums and subtractions (see Legendre & Legendre, 1998 and Muñoz *et al*., 2005).

A hierarchical partitioning procedure was also applied to the more relevant variables (Chevan & Sutherland, 1991; Mac Nally, 1996, 2000). This method aims to measuring the explanatory capacity of individual variables, considering all possible models ($2^k$, where $k$ is the number of variables considered) in a hierarchy and computing the additional explained deviance by adding any one variable to a simpler model that does not include that variable. Mean additional explained deviance per variable, $I_A$ (A denotes the given variable), is considered the explanatory power of each variable independently. The five climate variables and four land-use variables most significantly related with *M. calpeiana* presence/absence were hierarchical

partitioned in R (R Development Core Team, 2004). The hier.part package (Walsh & Mac Nally, 2003), restricted to factors monotonicly related with the dependent variable, was run for land-use variables, while climate factors, some of them related by quadratic functions, were manually hierarchical partitioned. Thus, the statistical significance of only the land-use variable $I_A$ could be calculated with 1000 randomizations of the matrix, followed by recompution of the distribution of $I_A$ (Mac Nally, 2002). Z ($I_{observed}$-$\bar{I}_{randomized}$/$SD_{Irandomized}$) is calculated and the statistical significance is based on the upper 95 percentile of the standard Normal distribution ($Z \geq 1.65$).

## RESULTS

*M. calpeiana* relationships with all the variables considered were statistically significant, except those with aspect and percentage of places without vegetation cover (Figs. 2 & 3). Annual precipitation, precipitation periodicity and annual temperature range are most explanatory, explaining 58.9, 58.4 and 49.1% of deviance, respectively; the first two positively, linearly related with spider presence-absence, and the third negatively related. NDVI, maximum summer temperature and minimum winter temperature are the three variables next in importance, with 35.3, 31.7 and 31.5 percentages of explained deviance, respectively. From a mid point, *M calpeiana* is positively, linearly related with NDVI. The relation with minimum winter temperature is positively linear until a threshold point (1-2ºC), where the slope of the curve approaches zero. The relation with maximum summer temperature is bell-shaped, with the maximum at ~28ºC. Slope and annual days of frost explain 17 and 16.1% of deviance, respectively. *M calpeiana* seems to avoid flat terrain, and prefers areas with

the smallest number of annual days of frost. Altitude, mean annual temperature and insolation explain negligible proportions of deviance (<4%); *M. calpeiana* avoids high altitudes (greater than 1500 m), low mean annual temperatures and medium insolation values. In general, the effect of land-use variables is low; most important are the percentage of forest and agricultural lands, with 8.56 and 12.8 percentages of explained deviance, respectively. The relationship with percentage of forest is complex, cubic, while the relationship with the percentage of agricultural land is negatively linear. Percentage of scrub and grassland, and the percentage of artificial surfaces, explain less than 3% of deviance, both positively, linearly related with *M. calpeiana* presence-absence (Fig. 3).

Correlation analysis identifies a group of highly-correlated explanatory variables, at the 0.8 threshold, composed of altitude, annual days of frost, mean annual temperature, and minimum winter temperature; the fourth variable, explaining significantly more variation than the other three, was selected as representative. Variables (and their transformations) selected for the subsequent final GLM analyses are: precipitation periodicity (linear); annual precipitation (linear); annual temperature range (linear); maximum summer temperature (quadratic); minimum winter temperature (cubic); slope (cubic); annual days of frost (linear); NDVI (cubic); percentage of forest (quadratic); percentage of agricultural land (linear); percentage of scrub and grassland (linear); and percentage of artificial surfaces (linear). Amount of forest was included as a quadratic term instead of cubic, as such a complex relationship with *M. calpeiana* presence would be of difficult biological interpretation. The inconsistent relation with insolation led to the elimination of this variable.

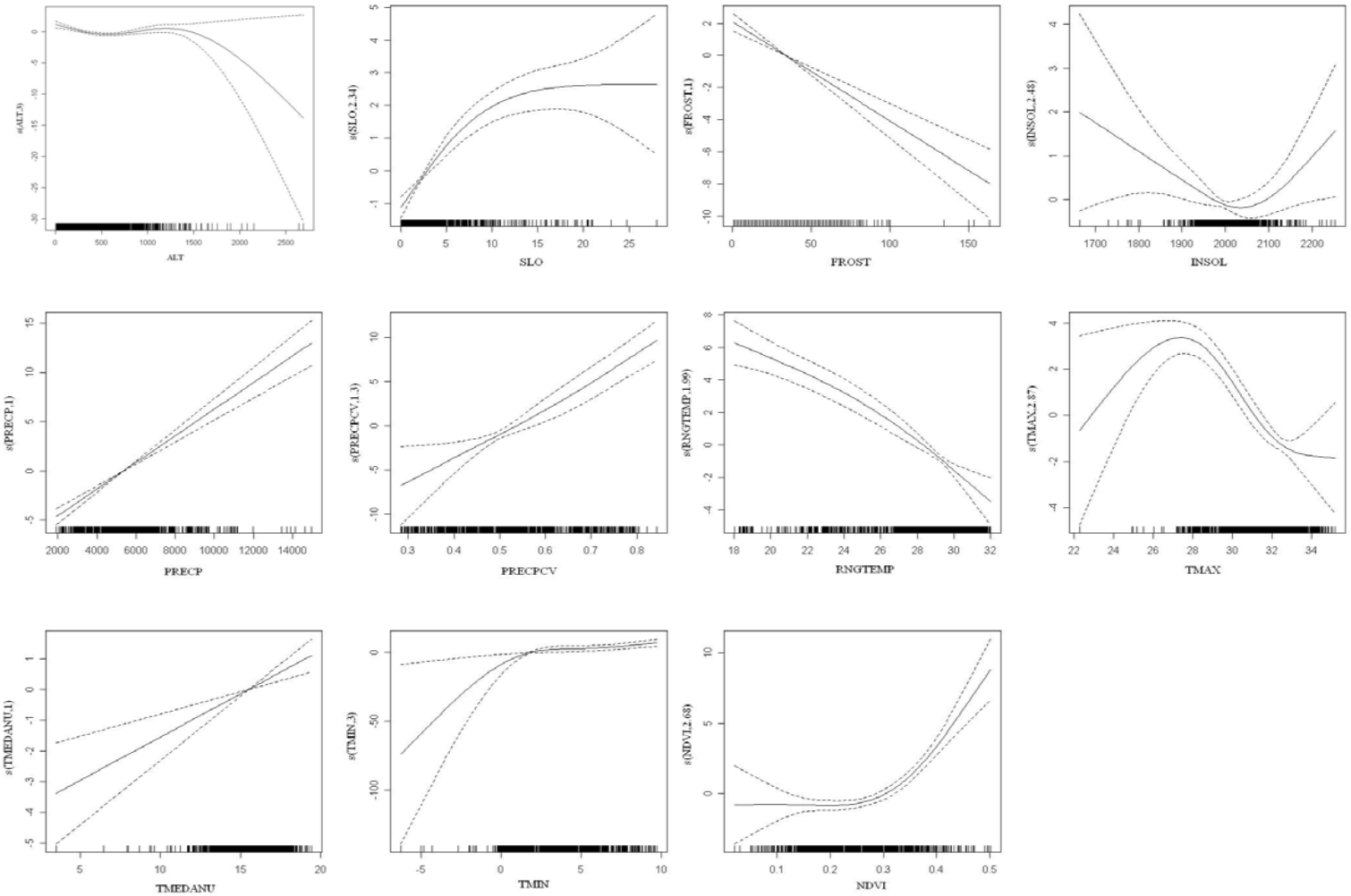



Figure 2.- Estimated GAM terms describing the relationships of *Macrothele calpeiana* (Walckenaer, 1805) with the statistically significant climate, topography and NDVI variables. Estimates are shown as solid lines, 95% confidence intervals as dashed ones and cases as a rough plot along graph bottom. Explained deviance: mean altitude ALT=3.68%, slope SLO=17%, annual days of frost FROST=16.1%, insolation INSOL=1.72%, annual precipitation PRECP=58.4%, precipitation periodicity PRECPCV=58.9%, annual temperature range RNGTEMP=49.1%, maximum annual temperature TMAX=31.7%, annual mean temperature TMEDANU=3.25%, minimum annual temperature TMIN=31.5%, Normalized Difference Vegetation Index NDVI=35.3%. All predictor *p*-values were lower than 0.01 (Chi. sq. test).

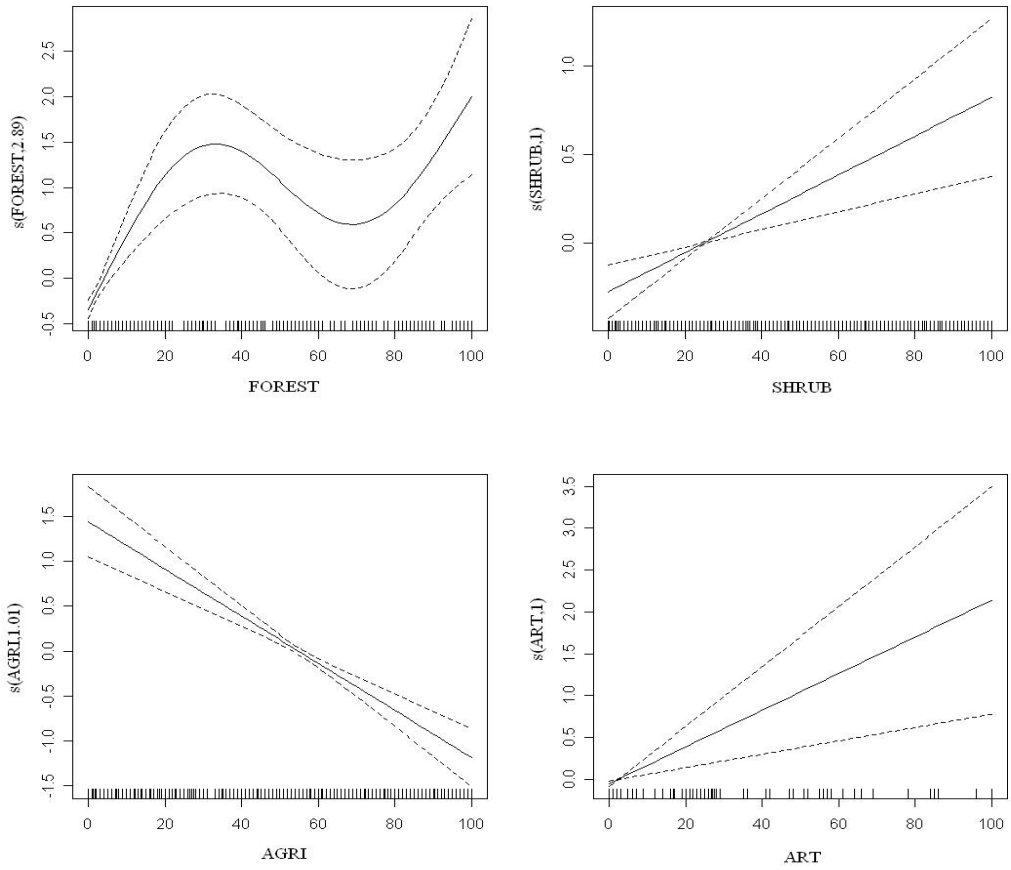



Figure 3.- Estimated GAM terms describing the relationships of *Macrothele calpeiana* (Walckenaer, 1805) with the statistically significant land-use variables. Estimates are shown as solid lines, 95% confidence intervals as dashed ones and cases as a rough plot along graph bottom. Explained deviance: percentage of woodland FOREST=8.56%; percentage of agricultural land AGRI=12.8%; percentage of scrub- and grassland SCRUB= 2.24%; percentage of artificial surfaces ART=1.43%. All predictor *p*-values were lower than 0.01 (Chi. sq. test).

The climate and topographic model retained only precipitation periodicity and annual precipitation as linear terms, and maximum summer temperature and minimum winter temperature as quadratic terms, while accounting for 82.15% of deviance and classifying almost perfectly (AUC = 0.99, sensitivity and specificity scores of 97%). This is the maximum predictive power achievable, since it is not increased by the inclusion of any other variable, neither NDVI nor land-use. A model based solely on NDVI accounts for 34.96% of deviance, with an AUC of 0.82 and sensitivity and specificity values of 76%. It includes the quadratic term of the NDVI variable. A model based only on land-use variables retains the linear terms of the percentage of

forest, agricultural lands and artificial surfaces, explaining only 14.76% of deviance (AUC = 0.76, sensitivity and specificity values of 70%). Lastly, a model combining both NDVI and land-use variables retained the three mentioned land-use variables and the quadratic term of NDVI, explaining 40.91% of deviance (AUC = 0.88, sensitivity and specificity scores of 0.79).

After application of the favourability function, final climate model interpolation to southern Spain (Fig. 4) highlighted the considerable favourability of mainly Huelva, Sevilla, Cádiz and Málaga provinces, corresponding to the principal *M. calpeiana* distribution areas. Suitable habitat extends also: through southern Granada; isolated potential areas on the coast of Valencia and Alicante; in eastern Jaén; and in the south-east of Cáceres province. Autocorrelation analysis showed that residuals of this final climate function are positive and significantly autocorrelated until a distance of 72 km (Fig. 5). From the added third degree polynomial of latitude and longitude (Legendre & Legendre, 1998), the former climate model retains only linear terms, raising explained deviance to 89.06% (almost a 7% increase), although the AUC values did not change (0.97) and sensitivity and specificity increased slightly, to 98%. Addition of spatial terms slightly decreases Moran's *I* autocorrelation scores for the first distance classes, although still positive and significant for the first six (Fig. 5).

Variation partitioning (Fig. 6) shows the importance of the effect of climate (41.2%) and the virtual lack of relevance of the independent effect of NDVI and land-use variables. The most important combined effect is that of climate and NDVI (26.1%), followed by the combined effect of the three groups of factors (8.8%) and the combined effect of climate and land-use (5.9%); the combined effect of land-use variables and NDVI did not explain any significant proportion of deviance.

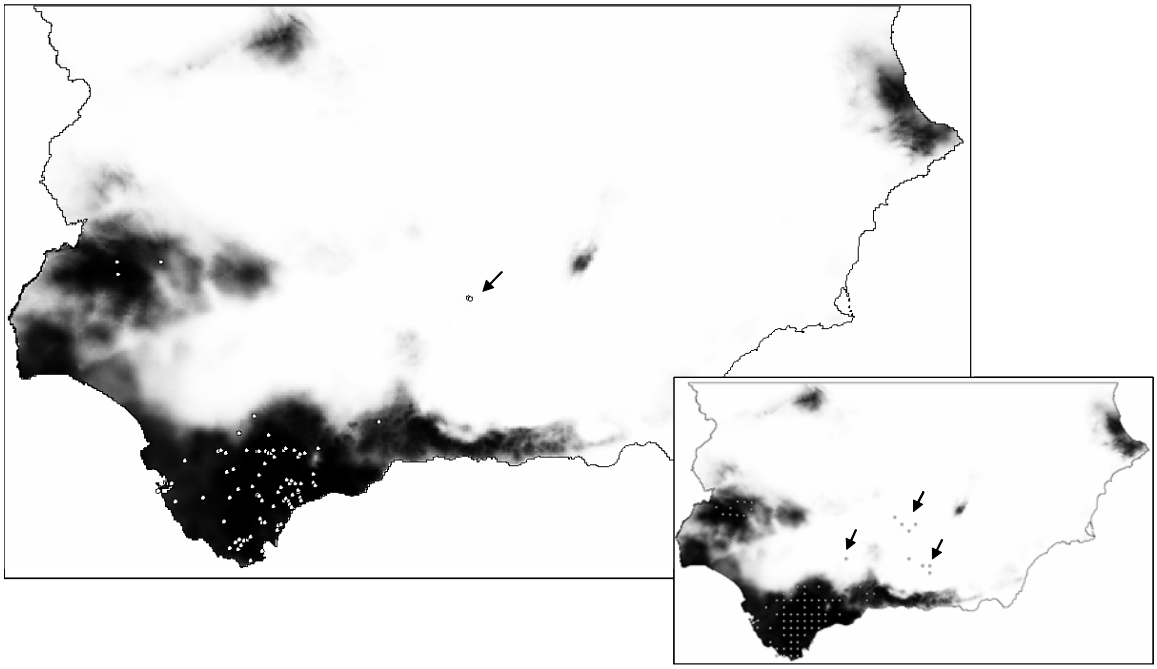


Figure 4.- Interpolated environmental favourability surface for *M. calpeiana* in southern Spain, with the 1×1 km presence points. In the small window, the 10×10 km presence points are overlaid; some core areas lack 1×1 km information. Arrows indicate occurrence areas not predicted by the climate model.

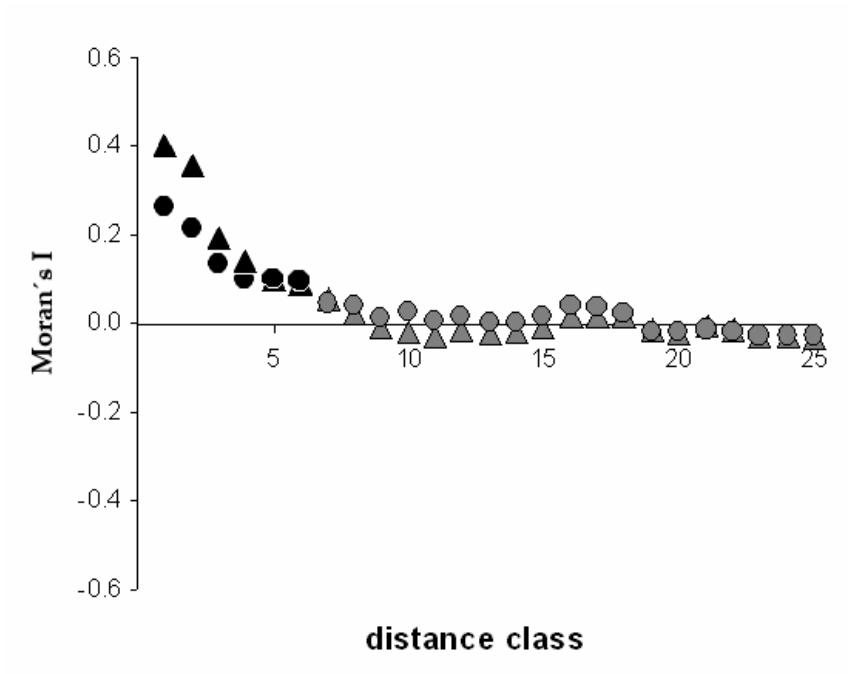



Figure 5.- Correlogram for the residuals of the model calculated after rescaling probabilities with the favourability function developed by Real *et al*. (in press) (see text for details). Triangles, climate model without spatial terms; circles, climate model with spatial terms. Lag distance is 12 kilometers and Moran's *I* autocorrelation scores were checked for significance (black dots are statistically significant) with a Bonferroni-corrected significance level (Sawada, 1999).

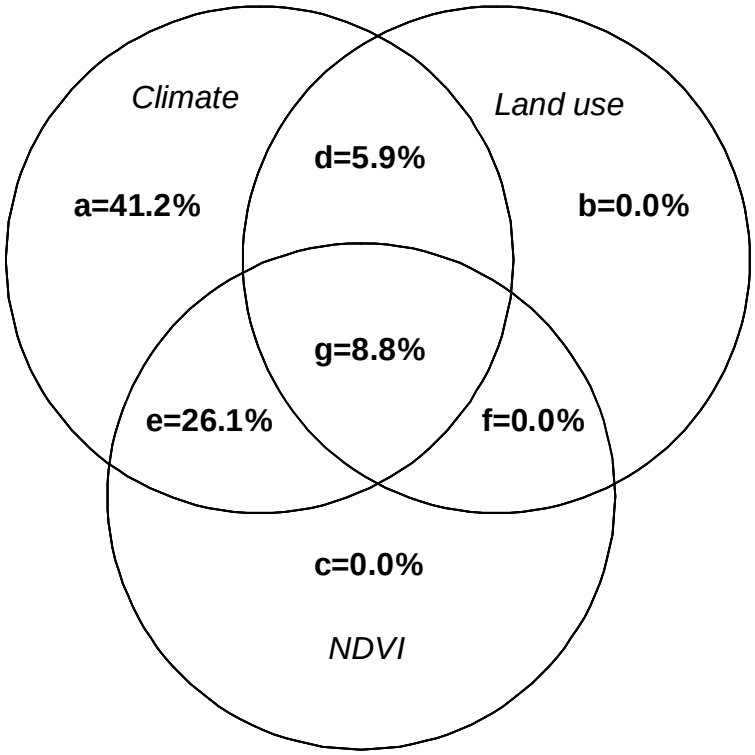



Figure 6.- Diagram of variation partitioning among climate, land use and NDVI groups of explanatory variables: a, b and c are independent effects of climate, land use, and NDVI, respectively; d, e and f are the combined variation due to the joint effect of climate and land use, the joint effect of climate and NDVI, and the joint effect of land use and NDVI, respectively; g is the combined variation due to the joint effect of the three groups of variables.

The independent effect of the most relevant climate variables was assessed using hierarchical partitioning (annual precipitation, precipitation periodicity, annual temperature range, maximum summer temperature and minimum winter temperature; see Fig. 7A). The greatest influence is exercised by annual precipitation and precipitation periodicity, with quite similar percentages of independent effects, 32.6% and 30.5%, respectively. Independent effects of the other three factors are smaller and similar (temperature range, 13.9%; maximum temperature, 12.2%; minimum temperature, 10.7%). Among the land-use variables (Fig. 7B), the greatest influence is exercised by percentage of agricultural land (51%), followed by percentage of forests (26.7%). Percentage of artificial surfaces and of scrub and grassland had smaller

percentages of independent effects, 12.0% and 10.3%, respectively. All these effects were statistically significant.

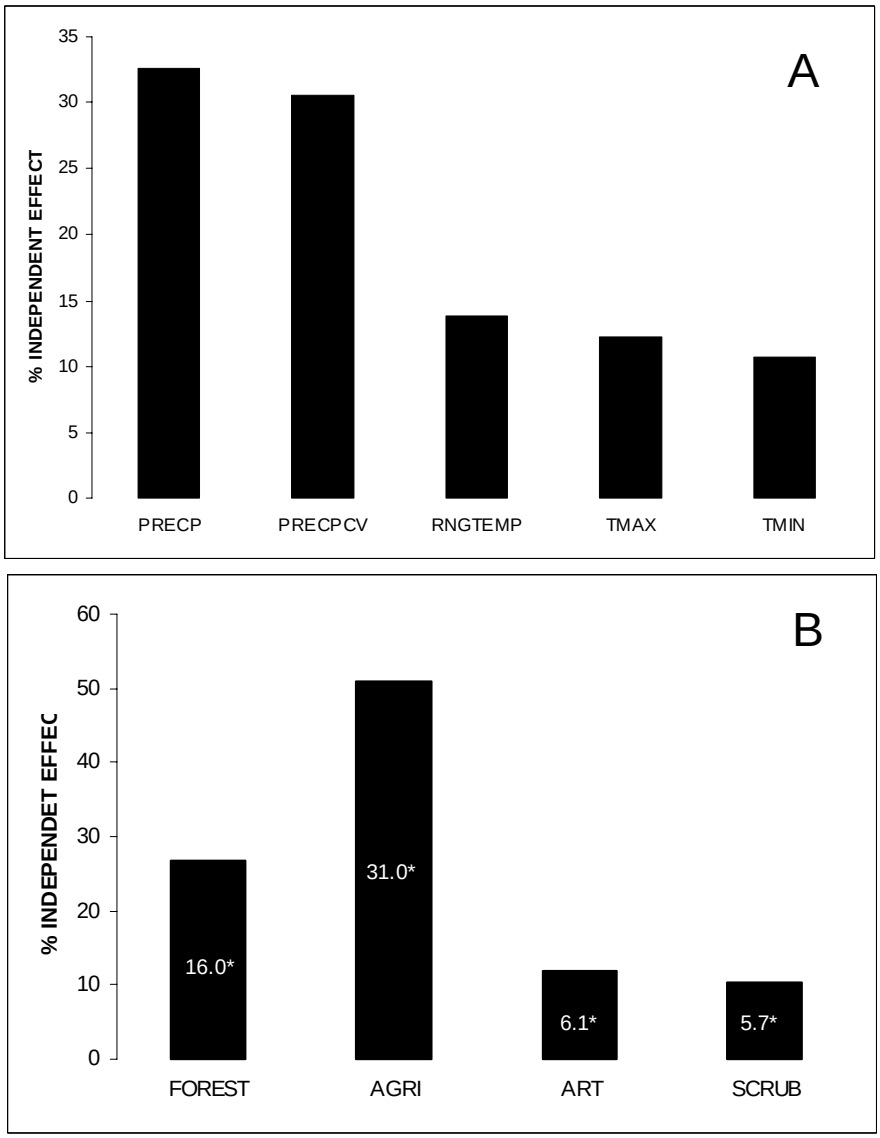



Figure 7.- Independent effects of climate (A) and land-use (B) variables calculated via hierarchical partitioning. Annual precipitation, PRECP; precipitation periodicity, PRECPCV; temperature range, RNGTEMP; maximum summer temperature, TMAX; minimum winter temperature, TMIN; percentage of woodland, FOREST; percentage of agricultural land, AGRI; percentage of artificial surfaces, ART; percentage of scrub- and grasslands, SCRUB. Numbers inside columns are Z-scores calculated using 1000 randomizations of the matrix, and statistically significant ones are marked (*).

# DISCUSSION

The distribution of *M. calpeiana* in the Iberian Peninsula, at a resolution of 1 $km^2$, is mainly determined by climate factors. Independent effects of precipitation-related variables, annual precipitation and precipitation periodicity, are similarly responsible for a great part (the greatest of the climate factors considered) of the variation in spider presence/absence. *M. calpeiana* lives in wet areas with high annual variation, clearly linearly related with these variables. Temperature-related variables are also important, though to a lesser extent than precipitation variables. While independent effects of temperature range, minimum and maximum temperature are similar, the combination of the last two variables is more important than temperature range. Minimum and maximum temperature remain in the final climate model, with curvilinear relations with the species, indicating a preference for places with moderate maximum temperatures and avoidance of areas with extreme minimum temperatures. Thus, the climate limitation of *M. calpeiana* distribution is evident even at the resolution of this study. The relevance of annual precipitation, precipitation periodicity and maximum and minimum temperature in determining the limits of *M. calpeiana* distribution is highlighted by the large single independent effect of climate (41.2%) in variation partitioning analysis. Moreover, the AUC value of the climate model (0.99) implies nearly perfect classification power (Swets, 1988).

Other climate and topography variables seem to be related with *M. capleiana*, although their low explicative power would indicate quite little relevance, at least on the scale of this study. In general, restricted by precipitation-related and maximum and minimum temperature variables, *M. calpeiana* shuns flat areas, prefers altitudes lower than 1500 m, and light, infrequent frosts, small temperature ranges and high

mean annual temperatures. In summary, *M. calpeiana* prefers thermo-mediterranean areas with maritime influence.

NDVI has been found by several authors as a good predictor of species distribution on a variety of scales (e.g., 100 $km^2$ resolution-global extent, Roura-Pascual *et al*., 2004; 1 $km^2$ resolution-regional extent, Osborne *et al*., 2001; Suárez-Seoane *et al*., 2002). At resolutions like that of our study, NDVI is only a complement to climate variables. In our study, NDVI did not remain in the model when added after climate variables. In fact, as shown by variation partitioning, its independent contribution was virtually null; its contribution to the variation in *M. calpeiana* presence/absence (34.96%) is inseparably correlated with that of climate and land-use variables. The AUC score, 0.82, of a model based only on NDVI was lower than that of the climate model, while 76% of both presences and absences were correctly predicted as such by the NDVI model. Thus, although it seems that *M. calpeiana* is positively related with degree of vegetative vigor, its effect cannot be separated from that of climate variables and, in any case, it is less relevant than climate.

Most land-use variables were weakly related with *M. calpeiana* presence/absence. We believe that possible land-use changes occurred since collection of the older records of the spider (the 1980s) may not have significantly altered possible relationships. The most important independent effect is that of agricultural land, which seems to impair species presence. The cubic relationship of percentage of woodland seems unrealistic, probably reflecting the effect of other constraining variables. In fact, when all land-use variables are included in a model, the percentage of forest remains as a linear term. So, *M. calpeiana* takes advantage of woodlands, the variable with the second-highest independent effect. A curious, statistically significant, positive relation with percentage of artificial surfaces appears, although its

independent effect is quite small. This pattern may reflect either the greater detectability of the spider in such areas, or the bias toward sampling in anthropogenized habitats, or *M. calpeiana* climate preferences for the conditions of the highly urbanized Iberian coast. Percentage of scrub and grassland, although positively related with spider presence, does not remain in the final land-use model, which explains a very low proportion of deviance (14.76%), with an AUC value of 0.76, and percentages of correctly predicted presences and absences of 70%. This AUC value is near the 0.7 value below which models should be regarded with scepticism, as in such cases sensitivity will not be much greater than the false positives fraction (Swets, 1988). Nevertheless, the slight effect of land-use variables cannot be separated from that of climate and NDVI variables, as shown by variation partitioning.

There have been claims, based on non-experimental observations of local density counts, that *M. calpeiana* is favoured by moderate human alteration of landscape (Helsdinge & Decae, 1992). Such favour is probably indirect, due to the creation of more potential nesting sites (Ferrández & Fernández de Céspedes, 1996). Our study, based on occurrence (not density) data, suggests that, on the study scale, agriculture is much more important than any other land-use type, affecting *M. calpeiana* presence negatively. The positive effect of woodlands is also much more relevant than the positive effect of percentage of artificial surfaces, although the anthropogenic impact advantageous for *M. calpeiana* may not be recognizable in our reclassification of Corine land-use classes into so few, broad categories. Also, our study scale may be too coarse to detect effects of land-use and moderate anthropogenic impact. Nevertheless, the relevance of natural cover in the distribution of *M. calpeiana* is independent of any local spider congregation in physical structures

that facilitate nesting. Habitat selection studies on finer scales than the present one must be carried out to develop firm conclusions about the impact of land-use on *M. calpeiana*. Additionally, once the species detectability factor has been accounted for (species may be more detectable in anthropogenized habitats), habitat suitability based on density data should be interpreted with caution. Abundance may vary in space due to a number of factors not related with long-term habitat favourability (see Van Horne, 1983 and Nielsen *et al*., 2005). In fact, attempts to model abundance data have generally failed to provide reliable models (e.g., Pearce & Ferrier, 2001; Nielsen *et al*., 2005). Thus, high density does not necessarily imply habitat suitability; reliable causal links can be obtained only from detailed demographic information (Van Horne, 1983; Mitchell, 2005).

A model run over all 14 significant variables, highlighting the importance of climate for *M. calpeiana,* retains only annual precipitation, precipitation periodicity and maximum and minimum temperatures. These four climate variables are the same determinants found in a 100-km$^2$-resolution study of the entire Iberian peninsula (see Jiménez-Valverde & Lobo, submitted). Thus, although causal explanations can not be consistently inferred from correlative analysis, the agreement between the results obtained on the two scales is suggestive. Interpolations to the southern Iberian peninsula of both models produce patterns of potential distribution that are chiefly coincident, but differing mainly in a reduction in potential area around the two core *M. calpeiana* distribution areas along both Guadalquivir river margins. Consequently, unsuitable area surrounding the Guadalquivir river basin is enlarged in the fine-scale model, which fails to predict two 1×1 km presence points located outside Jaén city (see arrow in Fig. 4). The model also fails to predict three core species areas where there were no fine-resolution occurrences (except the two in Jaén; compare maps of

Fig. 4): in the province of Granada; Jaén; the south of Córdoba. Precisely, these distribution-area environmental conditions are the most marginal where *M. calpeiana* can be found; their absence from the model training process may be cause of the general reduction in potential area and the observed underestimation. Representation of the full environmental and spatial gradient in the dependent variable is essential to obtain accurate models (Vaughan & Ormerod, 2003; Hortal & Lobo, 2005; Jiménez-Valverde & Lobo, 2006). Of special relevance, presence points at species environmental gradient limits define a distribution border in perhaps the most extreme conditions. Occurrence data not recovered from a well-designed sampling scheme, but from heterogeneous sources (bibliography, collections, etc.), may be biased (Dennis & Thomas, 2000; Jiménez-Valverde & Ortuño, in press). Although with sufficient information, false absences, i.e., failure to record focus-species occurences, can be detected (see, for example, Anderson, 2003) and so excluded from the modelling process; failure to include false absences as presences will nevertheless affect model results.

Apart from these data-dependent drawbacks, scale may be responsible for model differences. Scale differences (resolution and extent) affect variable relationships, and so too parameter estimations (Dungan *et al.*, 2002), highlighting the difficulty encountered in selection of the appropriate scale of analysis, which may not be straightforward. Moreover, most modelling is constrained by data availability, of both dependent and independent variables. Although the effects of scale are widely recognized (Wiens, 1989; Bailey *et al*., 2002; Pearson *et al*., 2004; Boyce, 2006) and multi-scale potential habitat studies are recommended (Martínez *et al*., 2003; Johnson *et al*., 2004; Luck, 2005; Oliver & Wotherspoon, 2005; Beever *et al*., 2006; Seoane *et*

*al.*, 2006), little is known about scale-of-analysis suitability to the structures and processes of study.

As in the case of other funnel-web spiders (see, for example, Woodman et al. 2006), *M. calpeiana*, a low-vagile species, is probably highly conditioned by local environmental factors. Thus, absence of local variables from the modelling process may be negatively affecting the rate of correct classification, as is corroborated by the slight decrease of autocorrelation in the first distance classes after the inclusion of spatial terms (Diniz-Filho *et al*., 2003).

## CONCLUDING REMARKS

In summary, climate is the main determinant of *M. calpeiana* distribution in southern Iberia at a 1×1 km resolution. In particular, precipitation-related variables are the most important factors for the species. On the scale of this work, no conclusive effect of land-use on *M. calpeiana* can be elucidated, although it may be suggested that preservation of natural vegetation is important for occurrence of the spider. The slight positive effect of artificial surfaces cannot be directly attributed to anthropogenized-habitat preferences.

We must stress the importance to distribution studies of detailed geo-referenced location data, as precise as possible to enable multi-scale approaches to habitat selection, to make reliable inferences about the process under study and to generate fully useful guidance for conservation proposes. Also, absence data is as important as presence data, or more so, to restrict predictions where needed. Thus, a measure of sampling effort should be reported for each sampling location in chorological studies, and absences reported as well as presences. Additionally, well-

designed field surveys must be carried out to recover all environmental and spatial variation of the target territory and avoid the use of biased data (Jiménez-Valverde & Lobo, 2004; Hortal & Lobo, 2005).

## AKCNOWLEDGEMENTS

Comments from Joaquín Hortal greatly improved the manuscript. This paper has been supported by an MEC Project (CGL2004-04309), as well as by a Museo Nacional de Ciencias Naturales/C.S.I.C./Comunidad de Madrid PhD grant.

## LITERATURE CITED

Anderson, R. P. (2003). Real vs. artefactual absences in species distributions: tests for *Oryzomys albigularis* (Rodentia: Muridae) in Venezuela. *Journal of Biogeography*, **30,** 591-605.

Bailey, S.-A., Haines-Young, R. H. & Watkins, C. (2002) Species presence in fragmented landscapes: modeling of species requirements at the national level. *Biological Conservation*, **108**, 307-316.

Beever, E. A., Swihart, R. K. & Bestelmeyer, B. T. (2006) Linking the concept of scale to studies of biological diversity: evolving approaches and tools. *Diversity and Distributions*, **12**, 229-235.

Boyce, M. S. (2006) Scale for resource selection functions. *Diversity and Distributions*, **12**, 269-276.

Buckland, S. T., Burnham, K. P. & Augustin, N. H. (1997) Model Selection: An Integral Part of Inference. *Biometrics*, **53**, 603-618.

Busby, J. R. (1991) BIOCLIM – A Bioclimate Analysis and Prediction System. In *Nature Conservation: Cost Effective Biological Surveys and Data Analysis*, eds. C. R. Margules & M. P. Austin, pp. 64-68. CSIRO, Australia.

Chevan, A. & Sutherland, M. (1991) Hierarchical partitioning. *American Statistician*, **45**, 90–96.

Chong, D. L. S., Mougin, E. & Gastellu-Etchegorry, J. P. (1993) Relating the global vegetation index to net primary productivity and actual evapotranspiration over Africa. *International Journal of Remote Sensing*, **14**, 1517- 1546.

Cramer, J. S. (1999) Predictive performance of binary logit model in unbalanced samples. *Journal of the Royal Statistical Society: Series D (The Statistican)*, **48**, 85-94.

Dennis, R. L. H. & Thomas, C. D. (2000) Bias in butterfly distribution maps: the influence of hot spots and recorder's home range. *Journal of Insect Conservation*, **4**, 73-77.

Diniz-Filho, J. A. F., Bini, L. M., Hawkins, B. A. (2003) Spatial autocorrelation and red herrings in geographical ecology. *Global Ecology & Biogeography*, **12**, 53-64.

Dungan, J. L., Perry, J. N., Dale, M. R. T., Legendre, P., Citron-Pousty, S., Fortin, M.-J., Jakomulska, A., Miriti, M. & Rosenberg, M. S. (2002) A balanced view of scale in spatial statistical analysis. *Ecography*, **25**, 626-640.

Engler, R., Guisan, A. & Rechsteiner, L. (2004) An improved approach for predicting the distribution of rare and endangered species from occurrence and pseudo-absence data. *Journal of Applied Ecology*, **41**, 263-274.

Ferrández, M. A. & Fernández de Céspedes, H. (1996) *Macrothele calpeiana (Walckenaer, 1805)*. In *Los Invertebrados no insectos de la "Directiva Hábitat" en España,* eds. M. A. Ramos, D. Bragado & J. Fernández, pp. 129-141. Dirección General de Conservación de la Naturaleza.

Fielding, A. H., & Bell, J. F. (1997) A review of methods for the assessment of prediction errors in conservation presence/absence models. *Environmental Conservation*, **24**, 38-49.

Gallon, R. C. (1994) Observations on *Macrothele calpeiana* (Walckenaer, 1805) in southern Iberia. *Journal of the British Tarantula Society Study Group*, **1**, 1-12.

Guisan, A. & Thuiller, W. (2005) Predicting species distributions: offering more than simple habitat models. *Ecology Letters*, **8**, 993-1009.

Guisan, A. & Zimmermann, N. E. (2000) Predictive habitat distribution models in ecology. *Ecological Modelling*, **135**, 147-186.

Harrell, F. E. J. (2001) *Regression modelling strategies: with applications to linear models, logistic regression, and survival analysis*. Springer, NY.

Helsdinge, P. J. van & Decae, A. E. (1992) Ecology, distribution and vulnerability of *Macrothele calpeiana* (Walckenaer) (Araneae, Hexathelidae). *Tijdschrift voor Entomologie*, **135**, 169-178.

Hortal, J. & Lobo, J. M. (2005) An ED-based protocol for optimal sampling of biodiversity. *Biodiversity and Conservation*, **14**, 2913-2947.

Jiménez-Valverde, A. & Lobo, J. M. (2004) Un método sencillo para seleccionar puntos de muestreo con el objeto de inventariar taxones hiperdiversos: el caso práctico de las familias *Araneidae* y *Thomisidae* (*Araneae*) en la Comunidad de Madrid, España. *Ecología*, **18**, 297-308.

Jiménez-Valverde, A. & Lobo, J. M. (2006) The ghost of unbalanced species distribution data in geographic model predictions. *Diversity and Distributions*, in press.

Jiménez-Valverde, A. & Ortuño, V. M. The history of endemic Iberian ground beetle description (Insecta, Coleoptera, Carabidae): which species were described first? *Acta Oecologica*, in press.

Jiménez-Valverde, A., Ortuño, V. M. & Lobo, J. M. Exploring the distribution of *Sterocorax* Ortuño, 1990 (Coleoptera, Carabidae) species in the Iberian Peninsula. *Journal of Biogeography*, in press.

Johnson, C. J., Seip, D. R. & Boyce, M. S. (2004) A quantitative approach to conservation planning: using resource selection functions to map the distribution of mountain caribou at multiple spatial scales. *Journal of Applied Ecology*, **41**, 238-251.

Legendre, P. & Legendre, L. (1998) *Numerical Ecology*. Elsevier, Amsterdam.

Lobo, J. M., Verdú, J. R. & Numa, C. (2006). Environmental and geographical factors affecting the Iberian distribution of flightless *Jekelius* species (Coleoptera: Geotrupidae). *Diversity and Distributions*, **12**, 179-188.

Luck, G. W. (2005) The habitat requirements of the rufous treecreeper (*Climacteris rufa*). 1. Preferential habitat use demonstrated at multiple spatial scale. *Biological Conservation*, **105**, 383-394.

Martínez, J. A., Serrano, D. & Zuberogoitia, I. (2003) Predictive models of habitat preferences for the Eurasian eagle owl *Bubo bubo*: a multiscale approach. *Ecography*, **26**, 21-28.

Mac Nally, R. (1996) Hierarchical partitioning as an interpretative tool in multivariate inference. *Australian Journal of Ecology*, **21**, 224–228.

MacNally, R. (2000) Regression and model-building in conservation biology, biogeography and ecology: the distinction between – and reconciliation of – "predictive" and "explanatory" models. *Biodiversity and Conservation*, **9**, 655–671.

Mac Nally, R. (2002) Multiple regression and inference in ecology and conservation biology: further comments on retention of independent variables. *Biodiversity and Conservation*, **11**, 1397-1401.

McCullagh, P. & Nelder, J. A. (1997) *Generalized Linear Models*. Monographs on Statistics and Applied Probability. Chapman & Hall, London.

Mitchell, S. C. (2005) How useful is the concept of habitat? – a critique. *Oikos*, **110**, 634-638.

Muñoz, A. R., Real., R., Barbosa, A. M. & Vargas, J. M. (2005) Modelling the distribution of Bonelli´s eagle in Spain: implications for conservation planning. *Diversity and Distributions*, **11**, 477-486.

Nielsen, S. E., Johnson, C. J., Heard, D. C. & Boyce, M. S. (2005) Can models of presence-absence be used to scale abundance? Two case studies considering extreme in life history. *Ecography*, **28**, 197-208.

Odland, J. (1988) *Spatial autocorrelation*. Sage Publications, Los Angeles.

Olden, J. D., Jackson, D. A. & Peres-Neto, P. (2002) Predictive models of fish species distributions: a note on proper validation and chance predictions. *Transactions of the American Fisheries Society*, **131**, 329-336.

Olivier, F. & Wotherspoon, S. J. (2005) GIS-based application of resource selection functions to the prediction of snow petrel distribution and abundance in East Antarctica: Comparing models at multiple scales. *Ecological Modelling*, **189**, 105-129.

Osborne, P. E., Alonso, J. C. & Bryant, R. G. (2001) Modelling landscape-scale habitat use using GIS and remote sensing: a case study with great bustards. *Journal of Applied Ecology*, **38**, 458-471.

Pearce, J. & Ferrier, S. (2001) The practical value of modeling relative abundance of species for regional conservation planning: a case study. *Biological Conservation*, **98**, 33-43.

Pearson, R. G., Dawson, T. P. & Liu, C. (2004) Modelling species distributions in Britain: a hierarchical integration of climate and land-cover data. *Ecography*, **27**, 285-298.

Perry, L. (2002) Captive breeding of the funnelweb spider *Macrothele calpeiana* (Walckenaer, 1805). *Journal of the British Tarantula Society*, **17(4)**, 113-121.

Pettorelli, N., Vik, J. O., Mysterud, A., Gaillard, J.-M., Tucker, C. J. & Stenseth, N. C. (2005) Using the satellite-derived NDVI to assess ecological responses to environmental change. *Trends in Ecology and Evolution*, **20**, 503-510.

R Development Core Team (2004) *R: A language and environment for statistical computing*. R Foundation for Statistical Computing, Vienna, Austria. Available at http://www.R-project.org

Real, R., Barbosa, A. M. & Vargas, J. M. Obtaining environmental favourability functions from logistic regression. *Environmental and Ecological Statistics*, in press.

Roura-Pascual, N., Suarez, A. V., Gómez, C., Pons, P., Touyama, Y., Wild, A. L. & Peterson, A. T. (2004) Geographical potential of Argentine ants (*Linepithema humile*

Mayr) in the face of global climate change. *Proceedings of the Royal Society of London B*, **271**, 2527-2535.

Santos Lobatón, M. C. (1996) Estudio sobre *Macrothele calpeiana* Walckenaer, 1805 (Araneae, Hexathelidae) en dos pinares de la provincia de Cádiz (España). *Aracnología*, **24**, 1-10.

Sawada, M. (1999) ROOKCASE: an Excel 97/2000 Visual Basic (VB) Add-in for exploring global and local spatial autocorrelation. *Bulletin of the Ecological Society of America*, **80**, 231–234.

Seoane, J., Justribó, J. H., García, F., Retamar, J., Rabadán, C. & Atienza, J. C. (2006) Habitat-suitability modelling to assess the effects of land-use changes on Dupont´s lark *Chersophilus duponti*: A case study in the Layna Important Bird Area. Biological Conservation, **128**, 241-252.

Silva, L. C. & Barroso, I. M. (2004) *Regresión logística*. La Muralla, Madrid.

Snazell, R. (1986) The spider genus *Macrothele* Ausserer in Spain (Araneae; Dipluridae). *Bulletin of the British Ecological Society*, **17**, 80-83.

Snazell, R. & Allison, R. (1989) The genus *Macrothele* Ausserer (Araneae, Hexathelidae) in Europe. *Bulletin of the British Arachnological Society*, **8(3)**, 65-72.

Suárez-Seoane, S., Osborme, P. E. & Alonso, J. C. (2002) Large-scale habitat selection by agricultural steppe birds in Spain: identifying species-habitat responses using generalizad additive models. *Journal of Applied Ecology*, **39**, 755-771.

Swets, J. A. (1988) Measuring the accuracy of diagnostic systems. *Science*, **240**, 1285-1293.

Van Horne, B. (1983) Density as a misleading indicator of habitat quality. *Journal of Wildlife Management*, **47**, 893-901.

Vaughan, I. P. & Ormerod, S. J. (2003) Improving the quality of distribution models for conservation by addressing shortcomings in the field collection of training data. *Conservation Biology*, **17**, 1601-1611.

Walsh, C. & Mac Nally, R. (2003) *hier.part: Hierarchical Partitioning.* R package version 0.5-1.

Wiens, J. A. (1989) Spatial scaling in ecology. *Functional Ecology* **3**: 385-397.

Wood, S.N. (2000) Modelling and smoothing parameter estimation with Multiple Quadratic Penalties. *Journal of the Royal Statistical Society Series B*, **62(2)**, 413-428.

Wood, S. N. (2004) *mgcv: GAMs with GCV smoothness estimation and GAMMs by REML/PQL.* R package version 1.1-8.

Wood, S. N. & Augustin, N. H. (2002) GAMs with integrated model selection using penalized regression splines and applications to environmental modelling. *Ecological Modelling*, **157**, 157-177.

Woodman, J. D., Ash, J. E. & Rowell, D. M. (2006) Population structure in a saproxylic funnelweb spider (Hexathelidae: *Hadronyche*) along a forested rainfall gradient. *Journal of Zoology*, **268**, 325-333.

Zaniewski, A. E., Lehmann, A. & Overton, J. McC. (2002) Predicting species spatial distributions using presence-only data: a case study of native New Zealand ferns. *Ecological Modelling*, **157**, 261-280.

**Annex.—** Distribution data sources (1 $km^2$ resolution) of *Macrothele calpeiana* in the Iberian Peninsula.

Blasco, A. & Ferrández, M. A. (1986) El género *Macrothele* Ausserer 1871 (Araneae; Dipluridae) en la Peninsula Ibérica. *Actas X Congr. Int. Aracnol. Jaca/España*, **I**, 311-320.

Helsdinge, P. J. van & Decae, A. E. (1992) Ecology, distribution and vulnerability of *Macrothele calpeiana* (Walckenaer) (Araneae, Hexathelidae). *Tijdschrift voor Entomologie*, **135**, 169-178.

Luque, F. J. R. (2001) Nuevos datos de *Macrothele calpeiana* (Walckenaer, 1805) para Jaén (España). *Revista Ibérica de Aracnología*, **4**, 34.

Santos Lobatón, M. C. (1996) Estudio sobre *Macrothele calpeiana* Walckenaer, 1805 (Araneae, Hexathelidae) en dos pinares de la provincia de Cádiz (España). *Aracnología*, **24**, 1-10.

Snazell, R. & Allison, R. (1989) The genus *Macrothele* Ausserer (Araneae, Hexathelidae) in Europe. *Bulletin of the British Arachnological Society*, **8(3)**, 65-72.

# Conclusiones y futuras líneas de trabajo

## ESTUDIO DE LA RIQUEZA ESPECÍFICA DE ARANEIDAE Y THOMISIDAE

I. Se pueden obtener inventarios fiables de las familias de arañas Araneidae y Thomisidae en parcelas de 1 $km^2$, siendo necesario el empleo complementario de tres técnicas de muestreo: mangueo, batido y trampas de caída. En localidades con una vegetación que propicie la concentración de araneidos en puntos claramente diferenciables en el paisaje, el uso de captura directa contribuye a mejorar el protocolo. Estos métodos han de combinarse en unidades de esfuerzo, cada una constando de 15 minutos de mangueo, 15 minutos de batido y 4 trampas de caída funcionando durante 48 horas. El uso de 20 unidades de esfuerzo de este tipo resulta suficiente para obtener inventarios rigurosos y poder realizar estimaciones de riqueza fiables.

II. Ante la imposibilidad de muestrear varias localidades durante todo un ciclo anual, los muestreos primaverales aportan una estima suficientemente completa de los ensamblajes, posibilitando la comparación de múltiples inventarios.

III. La inclusión de los individuos juveniles en las estimas de biodiversidad es recomendable siempre que sea posible, a fin de obtener mejores inventarios y mejores estimas de riqueza.

IV. El análisis de agrupamiento basado en el algoritmo *k-means* es una metodología aplicable a la selección de puntos de muestreo que, empleando variables ambientales y espaciales, permite considerar el gradiente espacio-ambiental del área de estudio. Esta técnica permite, además, maximizar dicha variabilidad ambiental y espacial en función del esfuerzo de muestreo que es posible desarrollar.

V. La riqueza local de arañas de las familias Araneidae y Thomisidae a escala regional está principalmente determinada por la complejidad estructural del hábitat. En concreto, la complejidad en los estratos herbáceos y subarbustivos llega a explicar el 81% de la variación en la riqueza específica.

A pesar del esfuerzo que se requiere para la realización del trabajo de campo, contar con inventarios fiables es esencial para abordar el estudio de la diversidad biológica. Es necesario establecer protocolos de inventariado rápido para el resto de familias de arañas, más teniendo en cuenta la acusada falta de datos corológicos con los que se cuenta en la Península Ibérica. También es necesario avanzar en la identificación taxonómica de los individuos juveniles, siendo prioritario ofrecer un mayor apoyo a la ciencia taxonómica. Es imprescindible implementar de una manera eficiente métodos de selección de puntos de muestreo (por ejemplo, el de la *p-median*) en programas de fácil manejo y libre acceso. Dada la elevada influencia de la estructura del hábitat en la determinación de la riqueza de arañas, sería útil estudiar la capacidad de otras variables disponibles en formato digital para dar cuenta de ella a fin de poder interpolar y extrapolar los resultados de los modelos a otras zonas del territorio. Por último, para comprender mejor la ecología del grupo y emplear estos conocimientos en conservación, sería fundamental corroborar el efecto de la complejidad del hábitat en la riqueza de otras familias de arañas y estimar si dicho efecto tiene o no una base directamente causal. Es decir, si se trata de una relación espuria, indirecta, o causal provocada por la capacidad de generación de refugios y lugares donde fijar las telas.

## ALGUNAS FUENTES DE ERROR EN LOS MODELOS PREDICTIVOS DE DISTRIBUCIÓN

I. Debido al efecto de la prevalencia sobre los valores generados por las probabilidades logísticas, carece de sentido seleccionar un punto de corte a 0.5 a fin de considerar la especie como presente. El punto de corte que minimiza la diferencia entre la sensitividad y la especificad es el que mejores predicciones proporciona. Emplear el punto de corte de 0.5 o el que maximiza el valor del Kappa puede implicar obtener estimas sesgadas de los errores de omisión y comisión.

II. Trabajar con prevalencias sesgadas no comporta mayores problemas que los que pueden solventarse reescalando las probabilidades obtenidas o aplicando un punto de corte adecuado con el que transformar esta variable continua en binaria. Sin embargo, conviene evitar valores de prevalencia menores de 0.01 y mayores de 0.99.

III. El tamaño de muestra de las presencias o las ausencias es una fuente importante de error en la realización de modelos predictivos de distribución, independientemente de sus tamaños relativos. Tamaños bajos de muestra implican una mala representación del gradiente ambiental. En el caso de disponer de pocas presencias y ausencias, los modelos resultantes sobreestimarán el rango de distribución. En estos casos, típicos cuando se trabaja con especies raras, aumentar el tamaño de muestra de las ausencias es esencial para restringir las predicciones, siendo deseable entonces una prevalencia sesgada de los datos.

IV. Las falsas ausencias son una fuente de error que, probablemente, esté siempre presente en los datos de distribución. Cuando estas ausencias se distribuyen siguiendo un patrón espacial, es imposible detectarlas y los modelos cometen, inevitablemente, errores de omisión. Ello significa infrapredecir el rango geográfico. Por el contrario, si las falsas ausencias se distribuyen al azar, las técnicas de modelización son capaces de corregir estos errores en la variable dependiente, aunque seguramente exista un umbral de proporción de falsas ausencias a partir del cual sea imposible parametrizar un modelo fiable.

Hasta ahora ningún estudio ha demostrado que exista una relación entre los valores obtenidos a partir de un método de modelización correlacional que utiliza datos de presencia/ausencia y las medidas directas de adecuación al hábitat basadas en estudios de campo, fisiológicos o demográficos. Establecer una relación directa, por ejemplo, entre las tasas de reproducción y supervivencia y los valores predichos por estos modelos es una cuestión prioritaria si queremos que las probabilidades obtenidas representen realmente *valores de adecuación*. Sin embargo, es probable que los modelos de distribución no sean en realidad modelos de *nicho* (*sensu* Hutchinson, 1957), sino meras aproximaciones a la distribución espacial de las especies. Estos modelos representarían hipótesis de distribución que estarían entre la distribución potencial y la real, sin incorporar estrictamente el nicho de las especies. Es por ello necesario establecer una base teórica sólida que clarifique cual es en realidad el objeto de estudio de las modelizaciones. Por otra parte, el que nos aproximemos más a la distribución real o a la potencial dependerá del tipo de datos usados y, más concretamente, del tipo de ausencias empleadas y del tipo de variables predictoras que

se utilicen. Estas dos cuestiones son básicas para comprender los resultados de las validaciones de los modelos, así como para interpretar correctamente las predicciones de forma que se elaboren hipótesis correctas. Es probable que la falta de este marco teórico que reivindicamos sea la causa de que en mucho trabajos no se establezca con claridad desde un comienzo qué se esta tratando de modelizar y, por tanto, cuál es el objetivo del estudio.

Existen multitud de técnicas con las que modelizar la distribución de las especies, y el desarrollo de nuevos y más potentes métodos capaces de parametrizar relaciones más complejas seguramente sea una interesante línea de trabajo (ver Elith *et al*., 2006). Sin embargo, aún poseemos un gran desconocimiento acerca de cómo funcionan las técnicas ya existentes y de cómo influyen las fuentes de incertidumbre asociadas a la variable dependiente en la fiabilidad de los modelos.

Una cuestión tan básica como es el criterio para seleccionar el punto de corte en una regresión logística no había sido abordada hasta la presente tesis doctoral. Somos conscientes de que unos mismos datos se pueden parametrizar con distintas formulas que producirán predicciones muy diferentes todas ellas con buenos valores de validación. Pero hasta que no comprendamos por qué las diferentes técnicas producen los resultados que producen seremos incapaces de generar hipótesis comprobables. Por tanto, sin renegar de los modelos de consenso o del desarrollo de nuevos métodos, creemos que es necesario un mayor esfuerzo para comprender el funcionamiento matemático de las técnicas ya existentes, proceso que ha sido evitado sutilmente por los biólogos dedicados a modelizar distribuciones (ver, por ejemplo, Pearson *et al.*, 2006).

Ahora que se hace práctica común emplear datos de distribución recopilados a partir de muy diversas y heterogéneas fuentes de origen, es más que nunca necesario comprender la influencia de los errores en la variable dependiente a modelizar. En este

sentido, más que los efectos puros, pueden ser especialmente interesantes las interacciones que se producen entre ellos. Entre otros temas, creemos necesario abordar el estudio:

- De la relación entre el ajuste del modelo y la prevalencia,
- De los efectos de la cantidad de ausencias y presencias en relación con el tipo de distribución de la especie (central, marginal),
- De los efectos de la relación entre la cantidad de falsos ceros y el tamaño de muestra,
- De los efectos de la distribución espacial y ambiental tanto de las presencias como de las ausencias, y su interacción con el tamaño de muestra

Es de sobra reconocido que el proceso de selección de hábitat está condicionado por múltiples factores que operan a diferentes escalas espaciales (resolución y extensión) y, por tanto, la escala afectará a los parámetros de los modelos y a su capacidad predictiva (Boyce, 2006). Más allá de la selección de hábitat, las predicciones de la distribución geográfica se ven igualmente afectadas por la escala. Sin embargo, a pesar de ser conscientes de su efecto, la elección de una u otra escala es aún un proceso cargado de subjetividad. Mientras que la resolución está condicionada, la mayoría de las veces, por la disponibilidad de los datos, tanto de los corológicos como de los predictores, la extensión suele determinarse en función de límites administrativos carentes de sentido biológico. Mientras que la resolución determinará la importancia relativa de las distintas variables, la extensión está íntimamente relacionada con el objetivo de la modelización, es decir, con la dualidad distribución real/potencial. Sin embargo, ambos conceptos, resolución y escala, condicionarán la extensión de la distribución predicha. El mayor problema es que la decisión de la escala idónea no

puede valorarse en función de los resultados de la validación de los modelos ya que, como se ha comprobado en esta tesis, con un mismo grupo de variables explicativas se pueden obtener modelos con altos valores de fiabilidad pero que, al interpolar, difieren en las distribuciones predichas. Cambios de resolución implican cambios en la media y la varianza de los valores de cada variable explicativa. Por otra parte, el área a partir de la cual muestreamos las ausencias, manteniendo constante su número, condicionará igualmente el rango de tolerancia de la especie, pareciéndose la distribución predicha más a la que obtendríamos a partir de un método que sólo emplea presencias cuanto mayor sea el área. La mayoría de los estudios sobre predicción de distribuciones muestran los resultados de análisis efectuados a una única escala. Es más que probable que los resultados se vieran alterados si la extensión se redujera o los datos se agruparan para trabajar a escalas menores. ¿Cómo podemos saber cual es la aproximación correcta? Dos trabajos efectuados con dos escalas distintas se publicarán por separado igualmente mientras se acompañen de los pertinentes análisis de validación para convencer a los revisores y editores de que los modelos desarrollados son "buenos", pero… ¿cómo podemos saber cuál se aproxima más a la realidad? Creemos que el problema de escala es uno de los mayores retos a los cuales se deberá enfrentar la modelización predictiva en los próximos años.

Todas estas cuestiones de índole metodológica pueden y deben ser abordadas mediante el empleo de especies virtuales, en las cuales se controlan todas las fuentes de error que operan en el mundo real y que inevitablemente oscurecen el verdadero efecto de la fuente de error que nos interesa. Es la única manera de experimentar en el campo de los modelos predictivos de distribución.

## ESTUDIO DE LA DISTRIBUCIÓN INDIVIDUAL DE *MACROTHELE CALPEIANA*

I. El régimen de precipitaciones parece ser el factor fundamental que determina la distribución de *M. calpeiana* en la Península Ibérica. La distribución potencial de la especie en la Península Ibérica es mayor que la real, siendo su ausencia de Portugal debida posiblemente a un efecto barrera del rió Guadiana.

II. La existencia de hábitat potencial en el Norte de África y la ausencia de registros de la especie para esta zona hace pensar en un origen y/o penetración oriental de la especie.

III. El cambio climático afectará negativamente a *M. calpeiana*, reduciendo su distribución actual conocida en la Península Ibérica. En Marruecos, su distribución potencial se verá fragmentado.

IV. A la escala de estudio utilizada no ha sido posible separar el efecto del uso del uso del de los factores climáticos. En cualquier caso, los resultados parecen indicar que la pérdida de hábitat natural y la extensión de las áreas de cultivo no han favorecido la presencia de *M. calpeiana.*

V. El escaso efecto positivo del suelo urbanizado puede deberse a relaciones no causales, como es la coincidencia de una elevada edificación en las áreas climáticamente más favorables para la araña.

De cara a elaborar una hipótesis sólida sobre la distribución y origen de *M. calpeiana* es necesario corroborar su ausencia en Portugal y Norte de África. Los muestreos deberían planificarse de tal manera que se recoja el máximo posible de la

variabilidad ambiental y espacial, debiendo proporcionarse alguna medida de esfuerzo de muestreo que permita estimar la fiabilidad de las ausencias obtenidas. Además, es necesario corroborar el estatus taxonómíco de las especies africanas de *Macrothele* y elaborar una hipótesis filogenética con las especies de este género, asi como estudios filogeográficos con las distintas poblaciones de *Macrothele calpeiana*.

*M. calpeiana* es una especie de interés en conservación, por lo que es esencial profundizar en el estudio de las variables que condicionan su distribución, densidad y procesos demográficos. Nuestros resultados muestran la existencia de factores no considerados que actuarían a una escala más fina de la empleada en esta tesis. Estudiar la selección de hábitat a escalas menores sería fundamental de cara a plantear estrategias de gestión y conservación para *M. calpeiana*. En este sentido, es importante disponer de buenas ausencias de campo con el fin de aproximarnos más a su distribución real. También seria necesario examinar si la actual red de espacios protegidos es capaz de representar las distintas poblaciones existentes de esta especie, análisis que debería hacerse extensivo a todo el conjunto de especies de invertebrados protegidos.

## REFERENCIAS BIBLIOGRÁFICAS

Boyce, M. S. (2006) Scale for resource selection functions. *Diversity and Distributions*, **12**, 269-276.

Elith, J., Graham, C. H., Anderson, R. P., Dudík, M., Ferrier, S., Guisan, A., Hijmans, R. J., Huettmann, F., Leathwick, J. R., Lehmann, A., Li, J., Lohmann, L.G., Loiselle, B. A., Manion, G., Moritz, C., Nakamura, M., Nakazawa, Y., Overton, J. M., Peterson, A. T., Phillips, S. J., Richardson, K., Scachetti-Pereira, R., Schapire, R. E., Soberón, J., Williams, S., Wisz, M. S. & Zimmermann, N. E. (2006) Novel methods improve prediction of species' distributions from occurrence data. *Ecography*, **29**, 129-151.

Hutchinson, G. E. (1957) Concluding remarks. *Cold Spring Harbor Symposia on Quantitative Biology*, **22**, 145-159.

Pearson, R. G., Thuiller, W., Araújo, M. B., Martínez-Meyer, E., Brotons, L., McClean, C., Miles, L., Segurado, P., Dawson, T. P. & Lees, D. C. (2006) Model-based uncertainty in species range prediction. *Journal of Biogeography*, en prensa.

# Anexos

**Anexo 1.- Listado de especies actualizado de las familias Araneidae y Thomisidae en la Comunidad de Madrid tras recopilar los datos existentes en la bibliografía, la colección del Museo Nacional de Ciencias Naturales, la colección privada del autor de la presente tesis, y añadir los datos obtenidos durante los muestreos de campo de la tesis doctoral (publicado en *Jiménez-Valverde, Lobo & López-Martos, Graellsia 2006, en prensa*).**

ARANEIDAE Latreille, 1806

Aculepeira Chamberlin & Ivie, 1942

*Aculepeira armida* (Audouin, 1826)

*Aculepeira ceropegia* (Walckenaer, 1802)

Agalenatea Archer, 1951

*Agalenatea redii* (Scopoli, 1763)

Araneus Clerck, 1758

*Araneus angulatus* Clerck, 1758

*Araneus diadematus* Clerck, 1758

*Araneus pallidus* (Olivier, 1789)

Araniella Chamberlin & Ivie, 1942

*Araniella alpica* (Koch, L., 1869)

*Araniella cucurbitina* (Clerck, 1758)

*Araniella inconspicua* (Simon, 1874)

*Araniella opisthographa* (Kulczynski, 1905)

Argiope Audouin, 1826

*Argiope bruennichi* (Scopoli, 1772)

*Argiope lobata* (Pallas, 1772)

Atea Koch, C.L., 1837

*Atea sturmi* (Hahn, 1831)

Cercidia Thorell, 1869

*Cercidia prominens* (Westring, 1851)

Cyclosa Menge, 1866

*Cyclosa algerica* Simon, 1885

*Cyclosa conica* (Pallas, 1772)

*Cyclosa insulana* (Costa, 1834)

Gibbaranea Archer, 1951

*Gibbaranea bituberculata* (Walckenaer, 1802)

*Gibbaranea gibbosa* (Walckenaer, 1802)

Hypsosinga Ausserer, 1871

*Hypsosinga albovittata* (Westring, 1851)

*Hypsosinga pygmaea* (Sundevall, 1831)

*Hypsosinga sanguinea* (Koch, C.L., 1844)

Larinioides Caporiacco, 1934

*Larinioides sclopetarius* (Clerck, 1758)

*Larinioides suspicax* (Pickard-Cambridge, O., 1876)

Mangora Pickard-Cambridge, O., 1889

*Mangora acalypha* (Walckenaer, 1802)

Neoscona Simon, 1864

*Neoscona adianta* (Walckenaer, 1802)

*Neoscona subfusca* (Koch, C.L., 1837)

Singa Koch, C.L., 1836

*Singa hamata* (Clerck, 1758)

Zilla Koch, C.L., 1834

*Zilla diodia* (Walckenaer, 1802)

Zygiella Pickard-Cambridge, F.O., 1902

*Zygiella x-notata* (Clerck, 1758)

THOMISIDAE Sundevall, 1833

Coriarachne Thorell, 1870

*Coriarachne* sp.

Firmicinus Simon, 1895

*Firmicinus bivittatus* Simon, 1895

Diaea Thorell, 1869

*Diaea dorsata* (Fabricius, 1777)

Heriaeus Simon, 1875

*Heriaeus mellotei* Simon, 1886

Misumena Latreille, 1804

*Misumena vatia* (Clerck, 1758)

Misumenops Pickard-Cambridge, F.O., 1900

*Misumenops* sp.

Ozyptila Simon, 1864

*Ozyptila atomaria* (Panzer, 1801)

*Ozyptila bicuspis* Simon, 1932

*Ozyptila blackwalli* Simon, 1875

*Ozyptila pauxilla* (Simon, 1870)

*Ozyptila perplexa* Simon, 1875

*Ozyptila umbraculorum* Simon, 1932

Pistius Simon, 1875

*Pistus truncatus* (Pallas, 1772)

Runcinia Simon, 1875

*Runcinia grammica* (Koch, C.L., 1837)

Synema Simon, 1864

*Synaema globosum* (Fabricius, 1775)

Thomisus Walckenaer, 1805

*Thomisus onustus* Walckenaer, 1805

Tmarus Simon, 1875

*Tmarus piochardi* (Simon, 1866)

*Tmarus staintoni* (Pickard-Cambridge, O., 1873)

*Tmarus stellio* Simon, 1875

Xysticus Koch, C.L., 1835

*Xysticus acerbus* Thorell, 1872

*Xysticus audax* (Schrank, 1803)

*Xysticus bifasciatus* Koch, C.L., 1837

*Xysticus bliteus* (Simon, 1875)

*Xysticus bufo* (Dufour, 1820)

*Xysticus caperatus* Simon, 1875

*Xysticus cor* Canestrini, 1873

*Xysticus cribatus* Simon, 1885

*Xysticus cristatus* (Clerck, 1758)

*Xysticus erraticus* (Blackwall, 1834)

*Xysticus ferrugineus* Menge, 1876

*Xysticus gallicus* Simon, 1875

*Xysticus grallator* Simon, 1932

*Xysticus kempeleni* Thorell, 1872

*Xysticus kochi* Thorell, 1872

*Xysticus lanio* Koch, C.L., 1835

*Xysticus lineatus* (Westring, 1851)

*Xysticus ninni* Thorell, 1872

*Xysticus nubilus* Simon, 1875

*Xysticus ovatus* Simon, 1876

*Xysticus robustus* (Hahn, 1832)

*Xysticus sabulosus* (Hahn, 1832)

*Xysticus semicarinatus* Simon, 1932

**Anexo 2.- Coordenadas UTM 10 × 10 km para cada especie de las familias Araneidae y Thomisidae en la Comunidad de Madrid (publicado en *Jiménez-Valverde, Lobo & López-Martos, Graellsia 2006, en prensa*).**

| **Araneidae** | |
|---|---|
| *Aculepeira armida* | 30TUK75 |
| | 30TUK97 |
| | 30TUK98 |
| | 30TVK09 |
| | 30TVK29 |
| | 30TVK43 |
| | 30TVK47 |
| | 30TVK49 |
| | 30TVK54 |
| | 30TVK66 |
| | 30TVL20 |
| | 30TVL31 |
| | 30TVL43 |
| | 30TVL52 |
| *Aculepeira ceropegia* | 30TUK98 |
| | 30TVK09 |
| | 30TVL10 |
| | 30TVL11 |
| | 30TVL23 |
| | 30TVL32 |
| *Agalenatea redii* | 30TUK86 |
| | 30TUK96 |
| | 30TVK09 |
| | 30TVK18 |
| | 30TVK26 |
| | 30TVK29 |
| | 30TVK38 |
| | 30TVK43 |
| | 30TVK48 |
| | 30TVK54 |
| | 30TVK56 |
| | 30TVK65 |
| | 30TVK75 |
| | 30TVL22 |
| | 30TVL30 |
| | 30TVL42 |
| | 30TVL43 |
| | 30TVL50 |
| | 30TVL52 |
| *Araneus angulatus* | 30TUK86 |
| | 30TUK98 |
| | 30TVK09 |
| | 30TVK49 |
| | 30TVL01 |
| | 30TVL10 |

| | |
|---|---|
| | 30TVL11 |
| | 30TVL31 |
| *Araneus diadematus* | 30TVK45 |
| | 30TVL11 |
| | 30TVL32 |
| *Araneus pallidus* | 30TUK98 |
| | 30TVK29 |
| | 30TVK36 |
| | 30TVK37 |
| | 30TVK43 |
| | 30TVK47 |
| | 30TVK48 |
| | 30TVK49 |
| | 30TVK65 |
| | 30TVL33 |
| | 30TVL41 |
| *Araniella alpica* | 30TVK09 |
| | 30TVL11 |
| | 30TVL12 |
| | 30TVL22 |
| *Araniella cucurbitina* | 30TUK98 |
| | 30TVK09 |
| | 30TVK18 |
| | 30TVK26 |
| | 30TVK38 |
| | 30TVK47 |
| | 30TVK48 |
| | 30TVK49 |
| | 30TVK65 |
| | 30TVK75 |
| | 30TVL01 |
| | 30TVL10 |
| | 30TVL11 |
| | 30TVL21 |
| | 30TVL22 |
| | 30TVL23 |
| | 30TVL30 |
| | 30TVL31 |
| | 30TVL32 |
| | 30TVL33 |
| | 30TVL41 |
| | 30TVL42 |
| | 30TVL43 |
| | 30TVL45 |
| | 30TVL52 |
| *Araniella inconspicua* | 30TVK29 |
| *Araniella opisthographa* | 30TVL11 |
| | 30TVL22 |
| | 30TVL30 |
| | 30TVL32 |

| | |
|---|---|
| | 30TVL45 |
| | 30TVL52 |
| *Argiope bruennichi* | 30TVK09 |
| | 30TVK16 |
| | 30TVK29 |
| | 30TVK36 |
| | 30TVK37 |
| | 30TVK38 |
| | 30TVK47 |
| | 30TVK48 |
| | 30TVK54 |
| | 30TVK57 |
| | 30TVK66 |
| | 30TVL42 |
| *Argiope lobata* | 30TUK75 |
| | 30TVK36 |
| | 30TVK47 |
| | 30TVK49 |
| | 30TVK54 |
| | 30TVK56 |
| | 30TVK57 |
| | 30TVK67 |
| *Atea sturmi* | 30TVL01 |
| *Cercidia prominens* | 30TVL32 |
| *Cyclosa algerica* | 30TUK86 |
| | 30TUK96 |
| | 30TUK98 |
| | 30TVK18 |
| | 30TVK38 |
| | 30TVK48 |
| | 30TVK49 |
| | 30TVK75 |
| | 30TVL30 |
| | 30TVL41 |
| | 30TVL45 |
| | 30TVL52 |
| *Cyclosa conica* | 30TUK86 |
| | 30TVK09 |
| | 30TVK29 |
| | 30TVK43 |
| | 30TVK47 |
| | 30TVL01 |
| | 30TVL11 |
| | 30TVL21 |
| | 30TVL22 |
| | 30TVL23 |
| | 30TVL50 |
| | 30TVL52 |
| *Cyclosa insulana* | 30TVK45 |
| | 30TVK56 |

| | |
|---|---|
| *Gibbaranea bituberculata* | 30TUK96 |
| | 30TVK09 |
| | 30TVK18 |
| | 30TVK29 |
| | 30TVK38 |
| | 30TVK43 |
| | 30TVK48 |
| | 30TVK54 |
| | 30TVK58 |
| | 30TVL41 |
| | 30TVL52 |
| *Gibbaranea gibbosa* | 30TUK96 |
| | 30TVK47 |
| | 30TVL01 |
| | 30TVL22 |
| *Hyposonga albovittata* | 30TUK96 |
| | 30TVK09 |
| | 30TVK18 |
| | 30TVK29 |
| | 30TVK48 |
| | 30TVK49 |
| | 30TVK54 |
| | 30TVK65 |
| | 30TVK78 |
| | 30TVL22 |
| | 30TVL30 |
| | 30TVL42 |
| | 30TVL52 |
| | 30TVK54 |
| | 30TVL32 |
| *Hyposinga pygmaea* | 30TVK45 |
| | 30TVL32 |
| *Hyposinga sanguinea* | 30TVK54 |
| *Larinioides sclopetarius* | 30TUK99 |
| | 30TVK09 |
| | 30TVK18 |
| | 30TVK29 |
| | 30TVK38 |
| | 30TVK47 |
| | 30TVK54 |
| | 30TVK57 |
| | 30TVK75 |
| | 30TVK86 |
| | 30TVL10 |
| | 30TVK19 |
| | 30TVL20 |
| | 30TVL21 |
| | 30TVL30 |
| | 30TVL41 |
| | 30TVL42 |

| | |
|---|---|
| | 30TVL52 |
| *Larinioides suspicax* | 30TVK38 |
| | 30TVL41 |
| *Mangora acalypha* | 30TUK86 |
| | 30TUK96 |
| | 30TUK98 |
| | 30TVK09 |
| | 30TVK18 |
| | 30TVK29 |
| | 30TVK37 |
| | 30TVK38 |
| | 30TVK43 |
| | 30TVK48 |
| | 30TVK49 |
| | 30TVK54 |
| | 30TVK65 |
| | 30TVK56 |
| | 30TVK75 |
| | 30TVL01 |
| | 30TVL20 |
| | 30TVL21 |
| | 30TVL22 |
| | 30TVL23 |
| | 30TVL30 |
| | 30TVL32 |
| | 30TVL41 |
| | 30TVL45 |
| | 30TVL52 |
| *Neoscana subfusca* | 30TUK96 |
| | 30TVK45 |
| | 30TVK65 |
| *Neoscona adianta* | 30TUK99 |
| | 30TVK29 |
| | 30TVK38 |
| | 30TVK48 |
| | 30TVK49 |
| | 30TVK54 |
| | 30TVK56 |
| | 30TVK64 |
| | 30TVK65 |
| | 30TVK19 |
| | 30TVL21 |
| | 30TVL22 |
| | 30TVL30 |
| | 30TVL31 |
| | 30TVL32 |
| | 30TVL41 |
| | 30TVL45 |
| | 30TVL52 |
| *Singa hamata* | 30TVK54 |

| | |
|---|---|
| *Zilla diodia* | 30TUK86 |
| | 30TUK96 |
| | 30TVK09 |
| | 30TVK29 |
| | 30TVK43 |
| | 30TVK48 |
| | 30TVK65 |
| | 30TVL23 |
| | 30TVL30 |
| | 30TVL32 |
| | 30TVL41 |
| | 30TVL45 |
| | 30TVL52 |
| | 30TVL01 |
| *Zygiella x-notata* | 30TVK65 |
| | 30TVL42 |
| **Thomisidae** | |
| *Coriarachne sp.* | 30TUK98 |
| *Diaea dorsata* | 30TVL22 |
| *Firmicinus bivittatus* | 30TVK09 |
| *Heriaeus mellotei* | 30TVK09 |
| | 30TVK48 |
| | 30TVK54 |
| | 30TVK65 |
| | 30TVL11 |
| | 30TVL32 |
| | 30TVL45 |
| *Misumena vatia* | 30TUK86 |
| | 30TUK96 |
| | 30TUK98 |
| | 30TVL01 |
| | 30TVL22 |
| | 30TVL45 |
| | 30TVK48 |
| *Misumenops sp.* | 30TVK09 |
| *Ozyptila atomaria* | 30TVL01 |
| | 30TVL22 |
| *Ozyptila bicuspis* | 30TVK48 |
| *Ozyptila blackwalli* | 30TVK48 |
| | 30TVK56 |
| *Ozyptila pauxilla* | 30TUK96 |
| | 30TVK29 |
| | 30TVK48 |
| | 30TVL52 |
| *Ozyptila perplexa* | 30TVK43 |
| *Ozyptila umbraculorum* | 30TVL45 |
| | 30TVL52 |
| *Pistus truncatus* | 30TUK96 |
| | 30TVK47 |
| | 30TVK48 |

| | |
|---|---|
| | 30TVL52 |
| *Runcinia grammica* | 30TUK86 |
| | 30TVK09 |
| | 30TVK47 |
| | 30TVK48 |
| | 30TVK54 |
| | 30TVK65 |
| | 30TVK67 |
| | 30TVK75 |
| | 30TVL30 |
| | 30TVL52 |
| *Synaema globosum* | 30TUK86 |
| | 30TUK96 |
| | 30TUK98 |
| | 30TVK09 |
| | 30TVK18 |
| | 30TVK29 |
| | 30TVK38 |
| | 30TVK43 |
| | 30TVK45 |
| | 30TVK47 |
| | 30TVK48 |
| | 30TVK54 |
| | 30TVK56 |
| | 30TVK65 |
| | 30TVK75 |
| | 30TVL01 |
| | 30TVL22 |
| | 30TVL30 |
| | 30TVL32 |
| | 30TVL45 |
| | 30TVL50 |
| | 30TVL52 |
| | 30TVL54 |
| *Thomisus onustus* | 30TUK86 |
| | 30TUK96 |
| | 30TUK98 |
| | 30TVK09 |
| | 30TVK18 |
| | 30TVK29 |
| | 30TVK37 |
| | 30TVK43 |
| | 30TVK47 |
| | 30TVK48 |
| | 30TVK54 |
| | 30TVK56 |
| | 30TVK65 |
| | 30TVK75 |
| | 30TVL01 |
| | 30TVL30 |

| | |
|---|---|
| | 30TVL45 |
| | 30TVL52 |
| *Tmarus piochardi* | 30TVK47 |
| *Tmarus staintoni* | 30TVK09 |
| | 30TVK37 |
| | 30TVK48 |
| | 30TVK65 |
| | 30TVL30 |
| *Tmarus stellio* | 30TVL45 |
| *Xysticus acerbus* | 30TUK96 |
| | 30TVK09 |
| | 30TVK18 |
| | 30TVK37 |
| | 30TVK48 |
| | 30TVL30 |
| *Xysticus audax* | 30TUK86 |
| | 30TVK29 |
| | 30TVK47 |
| | 30TVK65 |
| | 30TVK75 |
| | 30TVL01 |
| | 30TVL22 |
| *Xysticus bifasciatus* | 30TVK54 |
| *Xysticus bliteus* | 30TVK17 |
| | 30TVK29 |
| | 30TVK37 |
| | 30TVK38 |
| | 30TVK48 |
| | 30TVK54 |
| *Xysticus bufo* | 30TUK86 |
| | 30TVK09 |
| | 30TVK37 |
| | 30TVK38 |
| | 30TVK48 |
| | 30TVK56 |
| | 30TVK57 |
| | 30TVK65 |
| | 30TVK68 |
| | 30TVL11 |
| *Xysticus caperatus* | 30TUK98 |
| | 30TVK48 |
| *Xysticus cor* | 30TVK48 |
| | 30TVL11 |
| *Xysticus cribatus* | 30TVK56 |
| | 30TVL54 |
| *Xysticus cristatus* | 30TVL11 |
| | 30TVL45 |
| *Xysticus erraticus* | 30TVL01 |
| | 30TVL12 |
| | 30TVL22 |

| | |
|---|---|
| *Xysticus ferrugineus* | 30TUK86 |
| | 30TVK09 |
| | 30TVK18 |
| | 30TVK29 |
| | 30TVK37 |
| | 30TVK38 |
| | 30TVK48 |
| | 30TVL10 |
| | 30TVL22 |
| | 30TVL52 |
| *Xysticus gallicus* | 30TVL22 |
| *Xysticus grallator* | 30TVK47 |
| | 30TVK48 |
| *Xysticus kempeleni* | 30TVK09 |
| *Xysticus kochi* | 30TUK86 |
| | 30TUK96 |
| | 30TVK09 |
| | 30TVK29 |
| | 30TVK38 |
| | 30TVK48 |
| | 30TVK54 |
| | 30TVK86 |
| | 30TVL22 |
| | 30TVL30 |
| | 30TVL45 |
| | 30TVL52 |
| *Xysticus lanio* | 30TVK09 |
| | 30TVK43 |
| *Xysticus lineatus* | 30TVL11 |
| *Xysticus ninni* | 30TVL22 |
| | 30TVL45 |
| *Xysticus nubilus* | 30TUK96 |
| | 30TVK09 |
| | 30TVK18 |
| | 30TVK29 |
| | 30TVK37 |
| | 30TVK43 |
| | 30TVK47 |
| | 30TVK48 |
| | 30TVK54 |
| | 30TVK65 |
| | 30TVK75 |
| | 30TVL01 |
| | 30TVL52 |
| *Xysticus ovatus* | 30TVK56 |
| *Xysticus robustus* | 30TVL32 |
| *Xysticus sabulosus* | 30TVK09 |
| | 30TVK17 |
| | 30TVK38 |
| | 30TVK47 |

| | |
|---|---|
| | 30TVK48 |
| | 30TVK65 |
| | 30TVK77 |
| | 30TVL45 |
| *Xysticus semicarinatus* | 30TVL12 |
| *Xysticus ulmi* | 30TVK48 |

**Anexo 3.- Especies de Araneidae y Thomisidae colectadas durante los muestreos de la presente tesis doctoral, con el número de individuos por localidad.**

| | Cercedilla (30TVL0913) | La Acebeda (30TVL4750) | Garcisancho (30TVL2120) | Rascafría, Sillada de | Chapinería (30TUK9868) | Chinchón (30TVK5543) | Tielmes (39TVK7456) | El Escorial, Bosque de la Herrería (30TVK0291) | El Berrueco (30TVL5228) | Colmenar Viejo, Dehesa de Navalvillar (30TVL3506) | Hoyo de Manzanares (30TVK2295) | Madrid, Monte de Valdelatas (30TVK4287) | Perales de Tajuña (30TVK6654) | Pelayos de la Presa (30TUK8466) | Rascafría, Cerro Cardoso (30TVL2823) | Valdemorillo (30TVK1184) |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| *Aculepeira armida* | | | | | | 6 | | | 38 | | | | | | | |
| *Aculepeira sp.* | 2 | | | | 6 | | 8 | 4 | | 4 | 16 | | 3 | | | 35 |
| *Agalenatea redii* | | | | | 10 | 10 | 5 | 2 | 43 | 2 | 10 | 2 | 8 | 2 | 1 | 2 |
| *Araneus angulatus* | 1 | | | | | | | | | | | | | | | |
| *Araneus sp.* | 2 | 5 | 10 | | | | | | | | | | | | | |
| *Araniella alpica* | 22 | | 20 | | | | | 1 | | | | | | | | |
| *Araniella cucrbitina* | 8 | 19 | 33 | | | | 1 | 1 | | 1 | | 7 | | | 3 | |
| *Araniella inconspicua* | | | | | | | | | | | 2 | | | | | |
| *Araniella opisthographa* | | 2 | 2 | | | | | | 1 | 2 | | | | | | |
| *Araniella sp.* | | | | | 12 | | | | | | | | 4 | 18 | | 8 |
| *Atea sturmi* | 2 | | | | | | | | | | | | | | | |
| *Cyclosa algerica* | | 1 | | | | | 1 | | | 3 | | 5 | | 1 | | |
| *Cyclosa conica* | 19 | | 10 | | | | | | 6 | | | | | 1 | 1 | |
| *Cyclosa sp.* | | | | | | | | 10 | | | 1 | | | | | |
| *Diaea dorsata* | | | 7 | | | | | | | | | | | | | |
| *Gibbaranea bituberculata* | | | | | 1 | 1 | | 1 | 1 | | 4 | 4 | | | | 1 |
| *Gibbaranea gibbosa* | 2 | | 1 | | 2 | | | | | | | | | | | |
| *Gibbaranea sp.* | | | | | | | | | | | | | | 1 | | |
| *Heriaeus mellotei* | | 2 | | | | 2 | | 3 | | | | | | | | |
| *Heriaeus sp.* | | | | | | | | | 3 | | | 5 | | | | |
| *Hypsosinga alvovittata* | | | 3 | | 1 | 1 | | 1 | 4 | 1 | 7 | 1 | | | | 2 |

**Anexo 3.- Continuación.**

| | Cercedilla (30TVL0913) | La Acebeda (30TVL4750) | Rascafría, Sillada de Garcisancho (30TVL2120) | Chapinería (30TUK9868) | Chinchón (30TVK5543) | Tielmes (39TVK7456) | El Escorial, Bosque de la Herrería (30TVK0291) | El Berrueco (30TVL5228) | Colmenar Viejo, Dehesa de Navalvillar (30TVL3506) | Hoyo de Manzanares (30TVK2295) | Madrid, Monte de Valdelatas (30TVK4287) | Perales de Tajuña (30TVK6654) | Pelayos de la Presa (30TUK8466) | Rascafría, Cerro Cardoso (30TVL2823) | Valdemorillo (30TVK1184) |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| *Hypsosinga sanguinea* | | | | | 2 | | | | | | | | | | |
| *Mangora acalypha* | 2 | 14 | 3 | 314 | 45 | 96 | 54 | 152 | 24 | 185 | 91 | 115 | 45 | 4 | 12 |
| *Misumena vatia* | 2 | 2 | 3 | 1 | | | | | | | 6 | | 1 | | |
| *Misumenops sp.* | | | | | | | 1 | | | | | | | | |
| *Neoscona adianta* | | 4 | | | 11 | | | | | 22 | | | | | |
| *Ozyptila atomaria* | 3 | | 1 | | | | | | | | | | | | |
| *Ozyptila pauxilla* | | | | 2 | | | | 3 | | 5 | | | | | |
| *Ozyptila umbraculorum* | | 1 | | | | | | 1 | | | | | | | |
| *Ozyptilla sp.* | | | | | 8 | | | | 4 | | | | | | 1 |
| *Pistus truncatus* | | | | 5 | | | | 1 | | | 1 | | | | |
| *Runcinia grammica* | | | | | 96 | 18 | 1 | 2 | 9 | | 5 | 5 | | | |
| *Synaema globosum* | 2 | 9 | 4 | 52 | 7 | 15 | 33 | 55 | 9 | 96 | 51 | 17 | 23 | 1 | 16 |
| *Thomisus onustus* | 1 | 10 | | 3 | 52 | 2 | 5 | 8 | 16 | 10 | 16 | 9 | | | 4 |
| *Tmarus staintoni* | | | | | | | 1 | | | | 2 | 33 | | | |
| *Tmarus stellio* | | 1 | | | | | | | | | | | | | |
| *Tmarus sp.* | | | | 34 | | 7 | 6 | 4 | 2 | 2 | | | 12 | | |
| *Xysticus acerbus* | | | | 2 | | | | | 1 | | | | | | 6 |
| *Xysticus audax* | 14 | | 22 | | | 1 | | | | 2 | | 3 | 1 | 15 | |
| *Xysticus bliteus* | | | | | | | | | | 1 | | | | | |
| *Xysticus cristatus* | | 2 | | | | | | | | | | | | | |
| *Xysticus erraticus* | 5 | | 15 | | | | | | | | | | | 2 | |
| *Xysticus ferrugineus* | | | | | | | | 1 | | 2 | 2 | | | 1 | 1 |
| *Xysticus gallicus* | | | | | | | | | | | | | | 1 | |
| *Xysticus kempeleni* | | | | | | | 1 | | | | | | | | |
| *Xysticus kochi* | | 7 | | 2 | 6 | | 1 | 16 | 6 | 8 | 2 | | | 1 | |

**Anexo 3.- Continuación.**

| | **Cercedilla (30TVL0913)** | **La Acebeda (30TVL4750)** | **Rascafría, Sillada de Garcisancho (30TVL2120)** | **Chapinería (30TUK9868)** | **Chinchón (30TVK5543)** | **Tielmes (39TVK7456)** | **El Escorial, Bosque de la Herrería (30TVK0291)** | **El Berrueco (30TVL5228)** | **Colmenar Viejo, Dehesa de Navalvillar (30TVL3506)** | **Hoyo de Manzanares (30TVK2295)** | **Madrid, Monte de Valdelatas (30TVK4287)** | **Perales de Tajuña (30TVK6654)** | **Pelayos de la Presa (30TUK8466)** | **Rascafría, Cerro Cardoso (30TVL2823)** | **Valdemorillo (30TVK1184)** |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| *Xysticus lanio* | | | | | | | 4 | | | | | | | | |
| *Xysticus ninnii* | | 62 | | | | | | | | | | | | 1 | |
| *Xysticus nubilus* | 1 | | | 2 | 1 | 2 | 2 | | | 3 | 2 | | | | 14 |
| *Xysticus sabulosus* | | 1 | | | | | | | | | | | | | |
| *Xysticus sp.1* | 2 | | | 1 | | | | | | | | | | | |
| *Xysticus sp.2* | 6 | | | | | | | | | | | | | | |
| *Xysticus sp. 3* | | | 1 | | | | | | | | | | | | |
| *Xysticus sp. 4* | | | | | | | | 1 | | | | | | | |
| *Zilla diodia* | 3 | 3 | | 4 | | | 7 | 6 | 24 | 4 | 3 | 1 | 10 | | |
| *Zygiella sp.* | | | | | | | | 3 | | | | | | | |